

UNED

ciencias

La interpretación de los datos

Una introducción a la Estadística Aplicada



Alfonso García Pérez



Subido por:



Interfase IQ

Libros de Ingeniería Química y más



<https://www.facebook.com/pages/Interfase-IQ/146073555478947?ref=bookmarks>

**Si te gusta este libro y tienes la posibilidad,
cómpralo para apoyar al autor.**

La interpretación de los datos

Una introducción a la Estadística Aplicada

ALFONSO GARCÍA PÉREZ

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

LA INTERPRETACIÓN DE LOS DATOS. UNA INTRODUCCIÓN A LA ESTADÍSTICA APLICADA

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del Copyright, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamos públicos

© Universidad Nacional de Educación a Distancia
Madrid 2014

www.uned.es/publicaciones

© Alfonso García Pérez

ISBN electrónico: 978-84-362-6947-5

Edición digital: diciembre de 2014

*Dedicado a las personas
que ayudan a otras personas*

Prólogo

Este libro está pensado para lectores que no saben nada de Estadística y que quieren comenzar a entenderla. Fundamentalmente es un libro de conceptos pero la aplicación de los Métodos Estadísticos no sólo es el siguiente paso a dar, sino que ésta permitirá al lector una mejor comprensión de los conceptos.

Por esta razón, el libro está lleno de ejemplos. Aunque todos ellos se pueden resolver con la ayuda de una calculadora, es recomendable utilizar algún paquete estadístico para que el cálculo no interfiera en el aprendizaje de los conceptos.

Podrían utilizarse varios paquetes estadísticos aunque de entre ellos hemos preferido resolver los ejemplos con R, no sólo porque este paquete es gratuito y el más utilizado sino porque es el paquete estadístico que tiene una mayor proyección de futuro. Además, si más adelante el lector decide profundizar en el estudio de la Estadística Aplicada, con este software podrá ejecutar cualquier método estadístico que quiera con el mismo nivel de complejidad que el requerido en la aplicación de los Métodos Estadísticos elementales aquí estudiados.

Aunque en la bibliografía aparecen varias referencias para aprender a manejar R, en la dirección de la contraportada de este texto tiene una dirección de Internet en donde aparecen instrucciones para instalar R así como todos los comandos utilizados en la resolución de este libro. Simplemente con copiarlos y pegarlos en la línea de comandos de R obtendrá la misma solución que aparece aquí. También puede, lógicamente, teclear las instrucciones que acompañan la resolución de los ejemplos, pero no olvide que este libro es un libro de conceptos los cuales esperamos asimile fácilmente y le permitan abrir la puerta de la Estadística Aplicada, una materia cada día más necesaria.

Si efectivamente desea continuar profundizando en esta disciplina, una vez que hayan asimilado este texto, le recomendamos continuar con el libro *Estadística Aplicada: Conceptos Básicos* del mismo autor que éste.

Quiero terminar agradeciendo a Yolanda Cabrero la lectura detallada de una versión preliminar de este libro, la cual ayudó a mejorarlo.

Alfonso García Pérez
e-mail: `agar-per@ccia.uned.es`

Índice

1. Estadística Descriptiva	9
1.1. Introducción	9
1.2. Representaciones gráficas	9
1.2.1. Representaciones de datos de tipo cualitativo	10
1.2.2. Representaciones de datos de tipo cuantitativo	11
1.3. Medidas de posición	13
1.4. Medidas de dispersión	15
1.5. Distribuciones bidimensionales de frecuencias	19
1.5.1. Ajuste por mínimos cuadrados	21
1.5.2. Precisión del ajuste por mínimos cuadrados	25
2. Modelización y Estimación: La Distribución Normal	29
2.1. Introducción	29
2.2. La ley de Probabilidad Normal	31
2.3. La distribución t de Student	38
2.4. Estimación de la media poblacional	41
2.5. Estimación de la varianza poblacional: Distribución χ^2 de Pearson	43
2.6. Estimación del cociente de varianzas poblacionales: Distribu- ción F de Snedecor	44
3. Estimación por Intervalos de Confianza	47
3.1. Introducción	47
3.1.1. Cálculo de Intervalos de Confianza con R	49
3.2. Intervalo de confianza para la media de una población normal .	51
3.3. Intervalo de confianza para la media de una población no nece- sariamente normal. Muestras grandes	53
3.4. Intervalo de confianza para la varianza de una población normal	56
3.5. Intervalo de confianza para el cociente de varianzas de dos po- blaciones normales independientes	57
3.6. Intervalo de confianza para la diferencia de medias de dos po- blaciones normales independientes	59

3.7.	Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes	61
3.8.	Intervalos de confianza para datos apareados	63
4.	Contraste de Hipótesis	65
4.1.	Introducción y conceptos fundamentales	65
4.2.	Contraste de hipótesis relativas a la media de una población normal	73
4.3.	Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes	78
4.4.	Contraste de hipótesis relativas a la varianza de una población normal	82
4.5.	El contraste de los rangos signados de Wilcoxon	86
5.	Comparación de Poblaciones	91
5.1.	Introducción	91
5.2.	Análisis de la Normalidad	93
5.3.	Análisis de la Homocestacidad	95
5.4.	Transformaciones Box-Cox	98
5.5.	Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes	105
5.6.	Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes	111
5.7.	El contraste de Wilcoxon-Mann-Whitney	115
5.8.	Análisis de la Varianza	117
5.8.1.	Comparaciones Múltiples	120
5.9.	Contraste de Kruskal-Wallis	123
5.9.1.	Contraste χ^2 de homogeneidad de varias muestras . . .	125
6.	Modelos de Regresión	127
6.1.	Introducción	127
6.2.	Modelo de la Regresión Lineal Simple	128
6.3.	Análisis de los residuos	132
6.4.	Modelo de la Regresión Lineal Múltiple	133
6.5.	Otros Modelos Lineales	136
7.	Bibliografía	139

Capítulo 1

Estadística Descriptiva

1.1. Introducción

Los datos son el elemento más importante de la Estadística y, por tanto, su correcto tratamiento resulta esencial. En este capítulo veremos cómo representarlos, cómo resumirlos con una *medida de posición*, la media o la mediana y, finalmente, analizaremos lo concentrados que están los datos alrededor de la media con una *medida de dispersión*, la varianza o la desviación típica.

Estos tres aspectos, que analizaremos en las siguientes secciones, forman lo que se denomina *Estadística Descriptiva*. Primero consideraremos *datos unidimensionales* concluyendo el capítulo con el caso de *datos bidimensionales*, es decir, con el caso en el que los datos son el resultado de dos medidas unidimensionales en los individuos de la muestra tales como su Peso y su Talla, o su Edad y su Nivel de Educación, o su Sexo y su Sueldo Anual, porque los datos no son más que eso, el resultado de observar una o varias variables unidimensionales como la Talla, el Peso, etc., en los individuos que forman la *muestra*, entendida ésta como un grupo de individuos elegidos al azar de la población en estudio, población de la que deseamos obtener conclusiones mediante lo que se denomina *Inferencia Estadística*. De hecho, en Estadística el término población no sólo se refiere a un conjunto de personas sino al colectivo del que queremos sacar conclusiones.

Es decir, con la Estadística Descriptiva dejamos que los datos *hablen por sí mismos*, dándonos una *foto fija* de la población de la que queremos sacar conclusiones mediante la Inferencia Estadística.

1.2. Representaciones gráficas

Los datos unidimensionales son de dos clases: o bien proceden de la observación de una variable de tipo *Cualitativo*, como el Color del Pelo, o el Estado

Civil, variables cuyos “valores” no son numéricos: Rubio, Moreno, ..., en el primer caso, o Soltero, Casado, ..., en el segundo, o bien los datos proceden de una variable de tipo *Cuantitativo* como el Peso o la Talla que proporciona valores numéricos. La representación gráfica de los datos depende de la clase que éstos sean.

1.2.1. Representaciones de datos de tipo cualitativo

Los datos procedentes de observaciones de una variable de esta clase vendrán recogidos en una tabla en donde aparece el recuento de individuos que presentan los diferentes *valores* de la variable.

La representación gráfica habitual para este tipo de datos es el *Diagrama de Sectores* consistente en dividir un círculo en tantos sectores como *valores* tenga la variable cualitativa, asignando a cada sector un tamaño (ángulo) proporcional al número de individuos que presenten ese valor, número que se denomina *frecuencia absoluta del valor*.

Ejemplo 1.1

En un estudio sobre las razones por las que no fue completado un tratamiento de radiación seguido de cirugía en pacientes de cáncer de cabeza y cuello se obtuvieron los datos dados por la siguiente *distribución de frecuencias absolutas*,

<i>Causas</i>	<i>n_i</i>
Rehusaron cirugía	26
Rehusaron radiación	3
Empeoraron por una enfermedad ajena al cáncer	10
Otras causas	1
	40

Mediante una regla de tres se pueden determinar los ángulos que corresponden a los cuatro *valores* o clases de la variable Causas

Rehusaron cirugía:	234
Rehusaron radiación:	27
Empeoraron por una enfermedad ajena al cáncer:	90
Otras causas:	9

pero es más fácil obtener el Diagrama de Sectores con R ejecutando la secuencia de instrucciones

```
> x2<-c(26,3,10,1)
> pie(x2)
```

El problema es que, de esta forma, el ordenador elige unos colores arbitrarios y, lo que es más importante, denomina con simples números los sectores correspondientes a las clases que presenta la variable cualitativa. Si queremos que denomine de una manera concreta a

los sectores, debemos crear primero un vector de nombres, es decir, un vector de caracteres, como hacemos en (1), pudiendo crear también un vector de colores en (2), obteniendo el gráfico deseado al ejecutar (3)

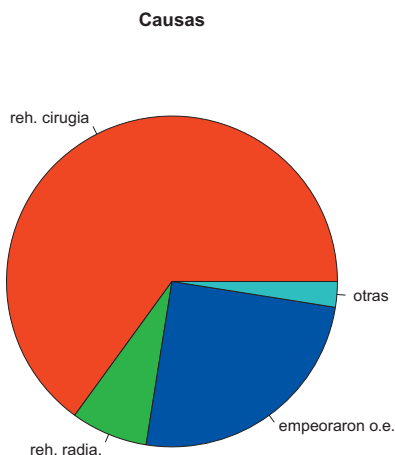


Figura 1.1 : Diagrama de Sectores del Ejemplo 1.1

```
> n2<-c("reh. cirugía","reh. radia.,"empeoraron o.e.,"otras")      (1)
> c2<-c(2,3,4,5)                                                    (2)
> pie(x2,labels=n2,col=c2)                                           (3)
```

El lector puede ir variando los números de los colores para obtener otro dibujo más de su agrado.

Si quisiéramos, además, poner título al gráfico podríamos utilizar otro argumento de la función `pie`, ejecutando (4), obteniendo finalmente, la Figura 1.1.

Apuntamos aquí que se denominan *funciones* de R a los programas incorporados a R cuya ejecución nos permitirá obtener determinados resultados. Estas funciones tienen *argumentos* u opciones para poder variar los resultados a obtener.

```
> pie(x2,labels=n2,col=c2,main="Causas")                            (4)
```

1.2.2. Representaciones de datos de tipo cuantitativo

En este caso los datos serán numéricos y la representación más habitual (aunque no la única) es el *Histograma* que consiste en una representación de

los datos en varios rectángulos cada uno de los cuales tiene un área (una altura si todos los rectángulos tienen la misma base) igual al número de individuos observados en dicho intervalo. Es posible elegir la amplitud de los intervalos (base de los rectángulos) en la representación, pero es más simple dejar que R lo haga.

Ejemplo 1.2

Se midieron los niveles de colinesterasa en un recuento de eritrocitos en $\mu\text{mol}/\text{min}/\text{ml}$ de 34 agricultores expuestos a insecticidas agrícolas, obteniéndose los siguientes datos:

Individuo	Nivel	Individuo	Nivel	Individuo	Nivel
1	10'6	13	12'2	25	11'8
2	12'5	14	10'8	26	12'7
3	11'1	15	16'5	27	11'4
4	9'2	16	15'0	28	9'3
5	11'5	17	10'3	29	8'6
6	9'9	18	12'4	30	8'5
7	11'9	19	9'1	31	10'1
8	11'6	20	7'8	32	12'4
9	14'9	21	11'3	33	11'1
10	12'5	22	12'3	34	10'2
11	12'5	23	9'7		
12	12'3	24	12'0		

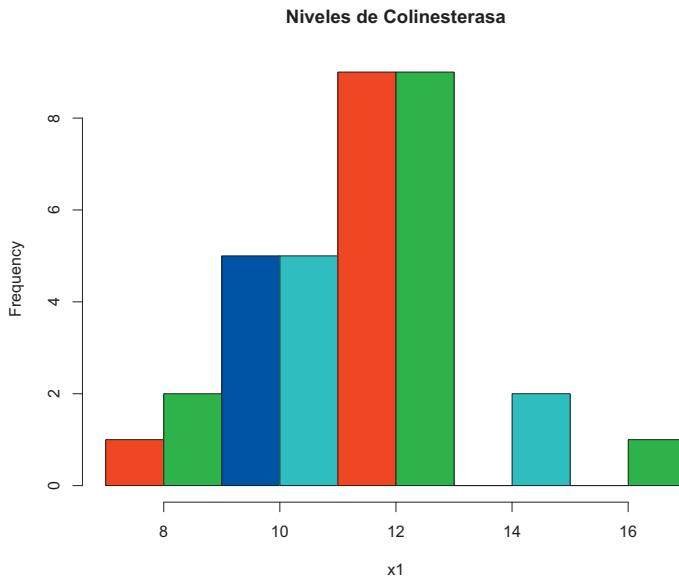


Figura 1.2 : Histograma del Ejemplo 1.2

Para representarlos, primero incorporamos los datos al ordenador y luego ejecutamos (1) obteniendo así el histograma, o ejecutamos (2) si queremos ponerle colores y título. De esta

última forma hemos obtenido la Figura 1.2.

```
> x1<-c(10.6,12.5,11.1,9.2,11.5,9.9,11.9,11.6,14.9,12.5,12.5,12.3,
+ 12.2,10.8,16.5,15,10.3,12.4,9.1,7.8,11.3,12.3,9.7,12,11.8,12.7,
+ 11.4,9.3, 8.6, 8.5, 10.1, 12.4, 11.1, 10.2)

> hist(x1) (1)

> hist(x1,col=c(2,3,4,5),main="Niveles de Colinesterasa") (2)
```

La otra representación gráfica que veremos para datos cuantitativos es el *Diagrama de hojas y ramas* que se obtendría ejecutando la función `stem` de R. Para el ejemplo anterior lo podemos obtener ejecutando

```
> stem(x1)

The decimal point is at the |

 7 | 8
 8 | 56
 9 | 12379
10 | 12368
11 | 11345689
12 | 0233445557
13 |
14 | 9
15 | 0
16 | 5
```

Como se ve, el diagrama de hojas y ramas es un histograma girado, con la misma interpretación visual que éstos, pero con una característica adicional muy importante: del gráfico podemos recuperar las observaciones; así, en este ejemplo, si empezamos a leer el gráfico por arriba, vemos que las observaciones son, 7'8, 8'5, 8'6, ..., 16'5.

1.3. Medidas de posición

En esta sección definiremos una serie de medidas o valores que representan o resumen un conjunto de datos, siendo también útiles, por tanto, para realizar comparaciones entre distintos grupos de datos. Estas medidas reciben el nombre de *promedios*, *medidas de posición* o *medidas de tendencia central* que, aunque alguna de ellas pueda aplicarse a caracteres cualitativos (como la Moda), habitualmente lo son sobre caracteres cuantitativos.

Media aritmética

La definición de *media aritmética* es simple. Se define como la suma de todos los valores observados dividido por el número de ellos. Más formalmente, como algunos de los valores observados pueden ser repetidos, si llamamos $x_1, \dots, x_k$ a los datos distintos de un carácter cuantitativo en estudio y $n_1, \dots, n_k$ a las correspondientes frecuencias absolutas de dichos valores, llamaremos media aritmética o simplemente media al valor

$$a = \frac{\sum_{i=1}^k x_i \cdot n_i}{n}$$

en donde el número total de observaciones n se denomina *frecuencia total*.

Ejemplo 1.2 (continuación)

Si sumamos todos los valores observados y dividimos por 34, la media aritmética o nivel medio de colinesterasa será,

$$a = \frac{10'6 + 12'5 + \dots + 10'2}{34} = 11'35$$

aunque es más fácil calcularlo con R ejecutando

```
> mean(x1)
[1] 11.35294
```

El [1] que sale antes del valor de la media es sólo para indicar el lugar de este valor y no debemos darle importancia.

Mediana

La otra medida de posición que estudiaremos es la *mediana* la cual se define como aquel valor de la variable tal que, supuestos ordenados los valores x_i de ésta en orden creciente, la mitad son menores o iguales y la otra mitad mayores o iguales. Así, si en la siguiente distribución de frecuencias absolutas

x_i	n_i
0	3
1	2
2	2
	7

ordenamos los valores en orden creciente,

0 0 0 1 1 2 2

el 1 será el valor que cumple la definición de mediana. No obstante, resulta más fácil calcularla con R mediante la función `median`.

```
> x3<-c(0,0,0,1,1,2,2)
> median(x3)
[1] 1
```

La mediana de los datos del Ejemplo 1.2, es decir, el nivel mediano de colinestarasa será

```
> median(x1)
[1] 11.45
```

La mediana es menos sensible a valores extremos de los datos puesto que por mucho que movamos el último dato (o el primero), la mediana seguirá siendo la misma.

Recordemos que la media de este conjunto de datos era 11'35. Cuando la media y la mediana de unos datos coinciden, se dicen que la distribución de frecuencias de estos datos es *simétrica* y en este ejemplo los datos muestran casi esa simetría, la cual se refleja en el histograma de la Figura 1.2.

1.4. Medidas de dispersión

Las medidas de posición estudiadas en la sección anterior servían para resumir los datos observados en un solo valor. Las *medidas de dispersión*, a las cuales dedicaremos esta sección, tienen como propósito estudiar lo concentrados que están los datos en torno a alguna medida de posición.

Estudiaremos sólo la *Varianza* y su raíz cuadrada, la *Desviación típica*.

Varianza

Si representamos por $x_1, \dots, x_k$ a los datos observados, llamaremos *Varianza* a la media aritmética de las desviaciones a la media a , es decir, a

$$s^2 = \frac{1}{n} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{1}{n} \sum_{i=1}^k x_i^2 n_i - a^2.$$

Al valor

$$S^2 = \frac{1}{n-1} \sum_{i=1}^k (x_i - a)^2 n_i = \frac{n s^2}{n-1}$$

se le denomina *cuasivarianza* y suele ser más utilizado que la propia varianza. De hecho, lo que R calcula con la función `var` es la cuasivarianza y será, por tanto, la medida habitual de dispersión que utilicemos.

Desviación típica

La varianza tiene un problema, y es que está expresada en unidades al cuadrado. Esto puede producir una falsa imagen de la dispersión de la distribución ya que no es lo mismo decir que la dispersión en torno a la estatura media es de 25 cm. que decir que es de 5 cm.; por esta razón suele utilizarse como medida de dispersión la raíz cuadrada de la varianza, denominada *Desviación típica*. Análogamente, la raíz cuadrada de S^2 se denomina cuasidesviación típica S y es calculada con la función `sd` de R.

Como, si el tamaño n de la muestra es grande, apenas hay diferencias entre la varianza y la cuasivarianza (y, por tanto, entre la desviación típica y la cuasidesviación típica), a veces se omite el prefijo cuasi para ambos valores aunque nosotros siempre los distinguiremos en el texto y hablaremos con precisión.

Ejemplo 1.2 (continuación)

La cuasivarianza y cuasidesviación típica de los niveles de colinesterasa antes utilizados son, respectivamente,

```
> var(x1)
[1] 3.514082
> sd(x1)
[1] 1.874588
```

Como vemos es más preciso decir que la dispersión de los datos es $1'87 \mu\text{mol}/\text{min}/\text{ml}$ que decir que es $3'51 \mu\text{mol}/\text{min}/\text{ml}$ al cuadrado.

Para finalizar esta sección trabajaremos un par de ejemplos aunque se recomienda al lector que se ejercite más con los libros de problemas resueltos que aparecen en la bibliografía del final del texto.

Ejemplo 1.3

Los tamaños (en hectáreas) de 25 asentamientos prehistóricos del Uruk tardío en la antigua Mesopotamia son, según Johnson (1973),

45	37	34'8	52	75	86	59'7	74	32	57'7
65	86	37	38'4	90'5	45	67	50	33	30
43'2	32	35'2	54'5	43'1					

Para hacer un Análisis Descriptivo de estos datos primero haremos una representación gráfica mediante un histograma ejecutando (2) después de introducir los datos con (1). El histograma obtenido aparecen en la Figura 1.3. Observamos que como hemos utilizado un vector con

cinco colores y tenemos siete intervalos, éstos se empiezan a repetir. Podemos modificarlo, si queremos, añadir o quitar colores.

```
> x<-c(45,37,34.8,52,75,86,59.7,74,32,57.7,65,86,
+ 37,38.4,90.5,45,67,50,33,30,43.2,32,35.2,54.5,43.1)
```

(1)

```
> hist(x,col=c1,main="Tamaño de asentamientos")
```

(2)

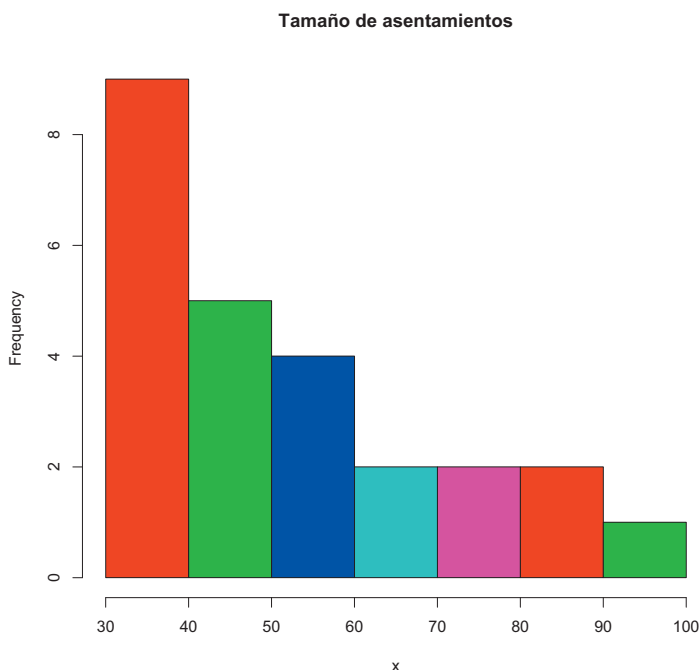


Figura 1.3 : Histograma del Ejemplo 1.3

Si quisiéramos que en el eje de ordenadas pusiera *Frecuencias absolutas* en lugar de *Frequency* teclearíamos

```
> hist(x,col=c3,main="Tamaño de asentamientos",ylab="Frecuencias absolutas")
```

utilizando un argumento más de la función `hist`. Análogamente se podría hacer con el eje la abscisas.

Ahora vamos a calcular algunas medidas de posición como la media (ejecutando (3)), la mediana (ejecutando (4)), y alguna medida de dispersión como la cuasivarianza (ejecutando (5)) y la cuasidesviación típica (ejecutando (6)).

```
> mean(x)
```

(3)

```
[1] 52.124
> median(x) (4)
```

```
[1] 45
> var(x) (5)
```

```
[1] 350.6494
> sd(x) (6)
```

```
[1] 18.3473
```

Se observa que la media y la mediana son bastante diferentes lo que indica una falta de simetría en los datos como de hecho se aprecia en el histograma de la Figura 1.3.

Ejemplo 1.4

Los siguientes datos corresponden al número de horas reales trabajadas en un año por 20 enfermeras de un determinado hospital, es decir, descontadas vacaciones, días de baja, etc. y añadidas las horas extras.

1235 , 1925 , 1850 , 1500 , 2015 , 1925 , 1750 , 1967 , 925 , 1500

1714 , 955 , 1800 , 1645 , 1992 , 1985 , 1555 , 1956 , 1962 , 2015

Si queremos hacer un Análisis descriptivo de estos datos, primero los incorporamos a R y después calculamos las medidas de posición y dispersión.

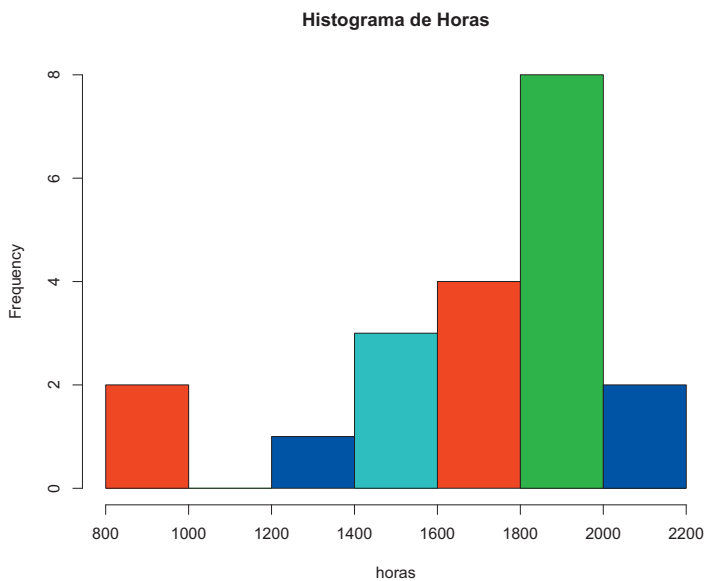


Figura 1.4 : Histograma del Ejemplo 1.4

```
> horas<-c(1235,1925,1850,1500,2015,1925,1750,1967,925,1500,1714,955,1800,  
+ 1645,1992,1985,1555,1956,1962,2015)  
  
> mean(horas)  
[1] 1708.55  
  
> median(horas)  
[1] 1825  
  
> var(horas)  
[1] 114806.2  
  
> sd(horas)  
[1] 338.8306
```

De nuevo se aprecia una fuerte asimetría en los datos y cómo la cuasidesviación típica es mucho más informativa en cuanto a la dispersión de los datos que la cuasivarianza. El histograma es el dado por la Figura 1.4 obtenido ejecutando

```
> hist(horas,main="Histograma de Horas",col=c(2,3,4,5))
```

1.5. Distribuciones bidimensionales de frecuencias

En esta sección estudiaremos la situación en la que los datos son observaciones de dos caracteres efectuadas en los individuos de una determinada muestra. Ambos caracteres pueden ser cuantitativos, como ocurre en el Ejemplo 1.5 de más abajo con el Peso y la Talla, pero también podrían ser ambos cualitativos, o uno cuantitativo y otro cualitativo. En todos estos casos los datos vendrán en forma de *tabla de doble entrada* en donde los valores de las dos variables definen las filas y las columnas, recogiendo en esa tabla el número de individuos de la muestra que presentan a la vez un valor y otro de ambas variables, como que entre los 80 individuos que forman la muestra del Ejemplo 1.5, hay 5 de Peso entre 70 y 80 kilos que además tienen una estatura entre 1'80 y 1'90 metros.

Ejemplo 1.5

Se observó el *Peso* y la *Talla* en 80 individuos, obteniéndose los siguientes datos,

	Talla	1'50 – 1'60	1'60 – 1'70	1'70 – 1'80	1'80 – 1'90	1'90 – 2'00
Peso						
50 – 60		2	1	1	2	2
60 – 70		3	3	2	4	8
70 – 80		5	4	3	5	4
80 – 90		2	4	2	6	6
90 – 100		1	2	1	5	2

En este libro, no obstante, nos vamos a centrar en el caso de que no existan pares de valores repetidos como ocurre en el Ejemplo 1.6 que sigue:

Ejemplo 1.6

Tras preguntar a 20 personas con aficiones atléticas la marca que poseían en 100 metros lisos y las horas semanales que por término medio dedicaban a entrenar, se obtuvieron los siguientes datos

Horas	21	32	15	40	27	18	26	50	33	51
Marca	13'2	12'6	13	12'2	15	14'8	14'8	12'2	13'6	12'6
Horas	36	16	19	22	16	39	56	29	45	25
Marca	13'1	14'9	13'9	13'2	15'1	14'1	13	13'5	12'7	14'2

Lo primero que analizamos es la representación gráfica de este tipo de datos. Para ello se utiliza el denominado diagrama de dispersión o *nube de puntos*, consistente en representar en un sistema de ejes coordenados de dos dimensiones tantos puntos como datos, asignando a cada dato (x_i, y_j) el punto de coordenadas (x_i, y_j) . La representación gráfica se obtiene utilizando la función `plot` de R.

Ejemplo 1.6 (continuación)

Para representar los datos, primero los incorporamos como indicamos en (1) y (2) y luego los representamos como decimos en (3). Se obtienen así la Figura 1.5. Aparecen después muchas posibles modificaciones del gráfico, invitando al lector a que los ejecute y a que los combine.

```
> x<-c(21,32,15,40,27,18,26,50,33,51,36,16,19,22,16,39,56,29,45,25) (1)
```

```
> y<-c(13.2,12.6,13,12.2,15,14.8,14.8,12.2,13.6,12.6,13.1,14.9,13.9,
+ 13.2,15.1,14.1,13,13.5,12.7,14.2) (2)
```

```
> plot(x,y) (3)
```

```
> plot(x,y,main="nube de puntos",col=3) # pone título y color los puntos
> plot(x,y,xlim=c(10,60),ylim=c(10,60)) # limita el recorrido del gráfico
> plot(x,y,pch="2") # pone los puntos como un 2
> plot(x,y,pch=2) # pone los puntos como el símbolo
```

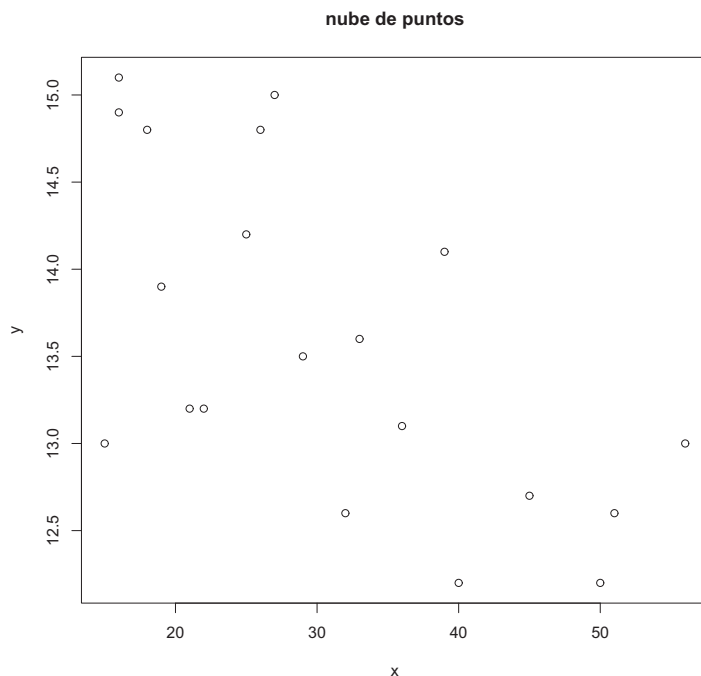


Figura 1.5 : Nube de puntos del Ejemplo 1.6

```

> plot(x,y,xlab="abscisa",ylab="ordenada") # número 2. Hay del 0 al 18
> plot(x,y,xlab=" ",ylab=" ")             # pone nombres a los ejes
> plot(x,y,axes=F)                         # no pone ningún nombre a los ejes
> plot(x,y,axes=F)                         # no pone el marco al gráfico

```

1.5.1. Ajuste por mínimos cuadrados

La Figura 1.5 parece mostrarnos gráficamente una idea razonable y es que, a medida que aumentemos el número de horas de entrenamiento, menor será la marca.

Lo mismo ocurre con el Peso y la Talla. Es un pensamiento común, la mayoría de las veces expresado de forma imprecisa, que el Peso y la Talla de los individuos de una población no son *independientes*, sino que por el contrario parece existir una determinada relación entre ellos, de forma que cuanto mayor sea la Talla de un individuo, mayor será su Peso.

La razón de tal idea se basa en la experiencia acumulada por las personas que ven una situación del tipo a la representada en la Figura 1.6, correspondiente a la nube de puntos del Peso y la Talla de 28 individuos.

Nos gustaría encontrar una fórmula que nos permitiera predecir el Peso y_i que obtendríamos para una Talla x_i determinada. En concreto nos gustaría determinar una recta que, sustituyendo en su fórmula

$$y_{t_i} = \beta_0 + \beta_1 x_i$$

una Talla determinada x_i , el valor *teórico* así obtenido y_{t_i} dado por la ecuación de esta recta, sea cercano al verdadero y_i .

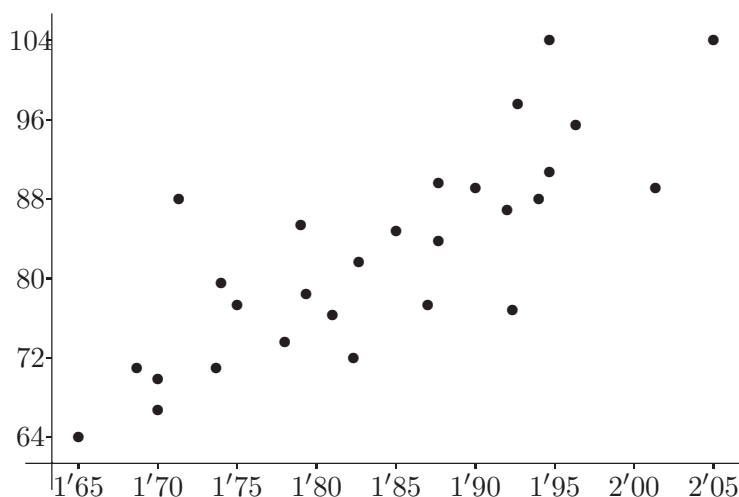


Figura 1.6

La recta que obtengamos así, es decir, determinando los valores β_0 y β_1 que minimicen las diferencias e_i entre los valores observados y_i y los teóricos y_{t_i} que nos dé esta recta, se denomina *recta de mínimos cuadrados*. Para evitar que esas diferencias se compensen entre positivas y negativas aunque sean muy grandes, se determina la recta más *próxima* a la nube de puntos (Figura 1.7), en el sentido de *mínimos cuadrados* de las diferencias, es decir, los valores de β_0 y β_1 que minimicen la suma de cuadrados

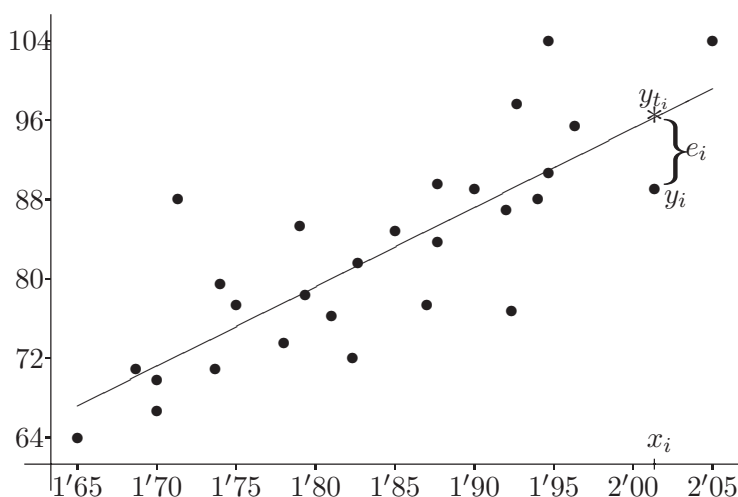


Figura 1.7

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - y_{t_i})^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Los valores así obtenidos son

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

y

$$\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$$

aunque la función `lm` de R hace los cálculos más rápido.

Esta recta de mínimos cuadrados se denomina también *recta de regresión* y los valores β_0 y β_1 , coeficientes de regresión (especialmente el segundo) aunque esta denominación tendrá su sentido en un contexto más amplio que estudiaremos más adelante en el que trataremos de explicar la *variable dependiente* Y en función de una (o más) *covariables independientes* X_i pero, de momento, es suficiente que sepamos que la recta antes determinada se puede denominar de ambas maneras.

Ejemplo 1.6 (continuación)

Si hiciésemos los cálculos mediante las fórmulas anteriores obtendríamos que la recta de mínimos cuadrados es

$$y = 15'05908 - 0'04786 x$$

cuya representación gráfica sobre la nube de puntos es la Figura 1.8, obtenida ejecutando la función `lm` como indicamos en (1). Dado que luego vamos a representarla sobre la nube de puntos, la asignamos un nombre, `ajus`, al ejecutar (1).

Si queremos ver cuál es la recta obtenida, ejecutamos (2), obteniendo en (3) la ordenada en el origen, 15'06, y la pendiente $-0'048$.

```
> ajus<-lm(y~x) (1)
```

```
> ajus (2)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x	
15.05908	-0.04786	(3)

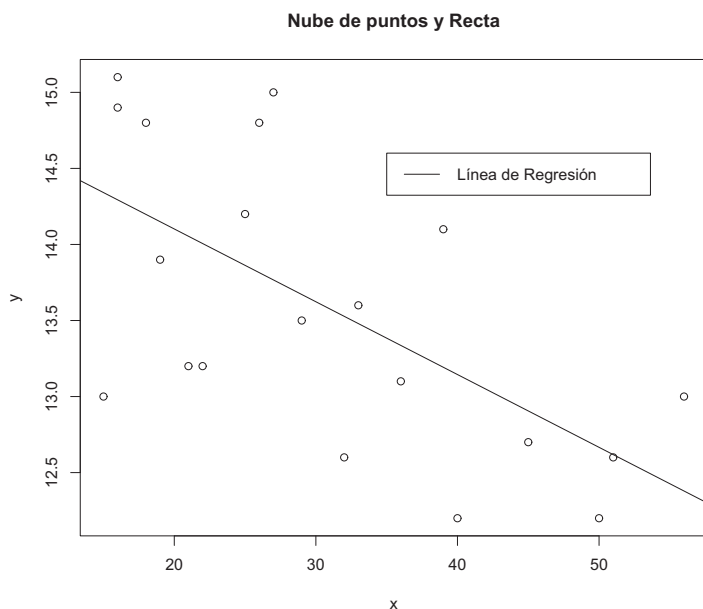


Figura 1.8: Nube de puntos y recta del Ejemplo 1.6

Podemos ahora añadirla a la nube de puntos, ponerle diferentes colores y diferentes grosores y, hasta poner un rótulo al gráfico, con las siguientes instrucciones. Invitamos al lector a ejecutarlas y combinarlas.

```
> abline(ajus) # añade la recta a la nube de puntos
> abline(ajus,col=2) # pone color a la recta de regresión
```

```
> abline(15.06,-0.048,lwd=2,col=4) # añade una recta de ordenada en el origen
                                     # 15.06, pendiente -0.048, grosor 2 y color 4
> legend(40,14.5,c("línea de regresión"),lty=c(1))
                                     # añade un rótulo en las coordenadas (40,14.5)
```

Destacamos cómo hemos podido añadir la recta simplemente dando su ordenada en el origen y su pendiente. Una posibilidad adicional es incluir una línea horizontal, `h`, en algún valor determinado `va1` de las ordenadas, y/o una línea vertical, `v`, en algún valor `va2` de las abscisas añadiendo a un gráfico ya existente la sentencia `abline(h=va1,v=va2)`; también se pueden poner colores. Nosotros hemos ejecutado la siguiente secuencia, además de (1), (2) y (3), para obtener la Figura 1.8,

```
> plot(x,y,main="Nube de puntos y Recta")
> abline(ajus,col=4)
> legend(35,14.6,c("Línea de Regresión"),lty=c(1),col=4)
```

1.5.2. Precisión del ajuste por mínimos cuadrados

La nube de puntos de la Figura 1.8 parece menos concentrada alrededor de su recta de ajuste que la recta de la Figura 1.7, lo que llevaría a pensar que la *predicción*

$$y = 15'05908 - 0'04786 \cdot 60 = 12'19$$

de la marca que obtendría un aficionado que entrenara 60 horas semanales no sería muy *fiable*.

La causa de esta falta de concentración de los valores observados alrededor de la recta puede ser que ambas variables no están relacionadas linealmente (un atleta nunca llegaría a hacer una marca negativa por muchas horas que se entrenase). Es posible que para este tipo de datos se ajustase *mejor* otro tipo de función.

Necesitamos, pues, un valor que nos dé una medida de lo próxima que está la función que hemos ajustado (sea o no una recta) a la nube de puntos de los datos; es decir, una medida de la *bondad del ajuste*. Este valor recibe el nombre de *Varianza Residual*

$$V_r = \frac{1}{n} \sum_{i=1}^n (y_i - y_{t_i})^2.$$

Aunque a la hora de comparar el ajuste de los datos por dos funciones podemos utilizar la varianza residual, siendo mejor aquella para la que dicha varianza sea menor, es conveniente utilizar otro valor que permita decidir si un ajuste es o no adecuado en sí mismo (puede que uno sea mejor que otro aunque ambos sean muy malos).

Surge así el concepto de *Coficiente de Determinación* definido como

$$R^2 = 1 - \frac{V_r}{s_y^2}$$

siendo V_r la varianza residual y $s_y^2 = \frac{1}{n} \sum_{i=1}^n (y_i - a_y)^2$ la varianza (marginal) de las y_i .

Este coeficiente está comprendido entre 0 y 1, hablándose de un buen ajuste en aquellos casos en los que R^2 esté cerca de 1, y de un mal ajuste en aquellos en los que sea cercano a 0. La valoración de lo que puede considerarse como *cerca* o *lejos*, deberá esperar hasta que aprendamos Inferencia Estadística.

Por último, veremos en esta sección un valor, relacionado con los anteriores en el caso de que se ajuste una recta. Se trata del *Coficiente de correlación lineal* de Pearson, definido como

$$r = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{\sqrt{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \sqrt{n \sum_{i=1}^n y_i^2 - (\sum_{i=1}^n y_i)^2}}$$

para el caso de que entre los n pares de datos no haya ninguno repetido.

Este coeficiente toma valores entre -1 y 1 , siendo $R^2 = (r)^2$ si se ha realizado el ajuste de una recta. La función `cor` de R calcula el valor de r .

Por último, digamos que para los datos del Ejemplo 1.6 el coeficiente de correlación es $r = -0'6304$

```
> cor(x,y)
[1] -0.6304069
```

y que, por tanto, el coeficiente de determinación es $R^2 = 0'3974$,

```
> cor(x,y)^2
[1] 0.3974129
```

Ejemplo 1.7

Los siguientes datos corresponden a un trabajo de Weiner(1977) en el que se midió el tamaño del vocabulario, es decir, el número de palabras que manejaban niños de diversas edades.

Edad	1	1'5	2	2'5	3	3'5	4	4'5	5	6
N. palabras	3	22	272	446	896	1222	1540	1870	2072	2562

Vamos a determinar la recta de regresión del Número de palabras en función de la Edad,

$$\text{Número de palabras} = \beta_0 + \beta_1 \text{ Edad.}$$

Para ello ejecutamos la siguiente secuencia de instrucciones

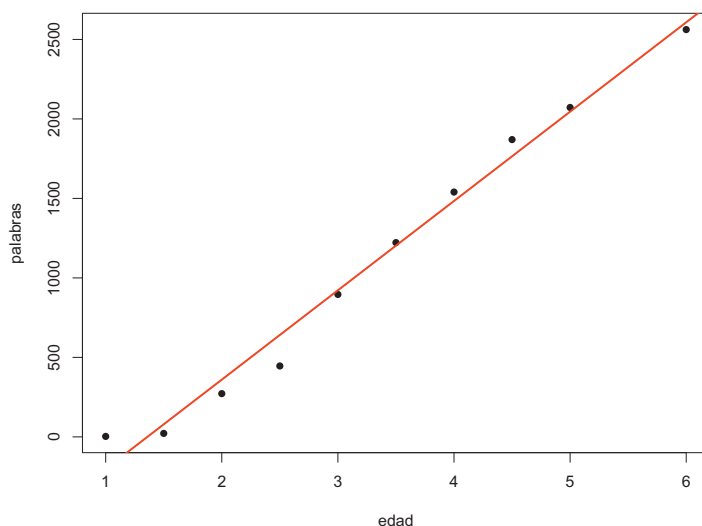


Figura 1.6 : Diagrama de dispersión y recta de regresión

```
> edad<-c(1,1.5,2,2.5,3,3.5,4,4.5,5,6)
> palabras<-c(3,22,272,446,896,1222,1540,1870,2072,2562)
> recta<-lm(palabras~edad)
> recta
Call:
lm(formula = palabras ~ edad)
```

```
Coefficients:
(Intercept)      edad
    -763.9       561.9
```

obteniendo que la recta de regresión es la de ecuación

$$\text{Número de palabras} = -763'9 + 561'9 \text{ Edad.}$$

Ejecutando la siguiente secuencia obtenemos la Figura 1.6 correspondiente a la nube de puntos y la recta de regresión calculada sobre ella.

```
> plot(edad,palabras,pch=16)
> abline(recta,col=2,lwd=2)
```

Para analizar la bondad del ajuste ejecutamos

```
> cor(edad,palabras)^2  
[1] 0.985272
```

valor que parece indicar un buen ajuste ya que la recta determinada permite *explicar* el Número de palabras mediante la Edad con un 98'5 % de fiabilidad.

Capítulo 2

Modelización y Estimación: La Distribución Normal

2.1. Introducción

En el capítulo anterior estudiamos cómo podemos representar y resumir unos datos. Habitualmente estos datos serán una muestra extraída de una población de la que queremos obtener conclusiones mediante un proceso que denominaremos *Inferencia Estadística* y al que dedicaremos el resto del libro. El término *población* no siempre se referirá a un conjunto de personas sino que lo entenderemos como el colectivo del que queremos obtener conclusiones.

Así por ejemplo, los 34 agricultores del Ejemplo 1.2 serán una muestra representativa de los agricultores expuestos a insecticidas agrícolas, grupo del que queremos obtener conclusiones como conocer (estimar) cuál es su nivel medio de colinesterasa, es decir, la media de la población, ya que este valor, denominado *parámetro*, permitirá valorar la magnitud de la contaminación.

El adjetivo *representativa* es muy importante para una muestra ya que es su propiedad clave. Si una muestra no fuera representativa, no podríamos sacar conclusiones de la población de la que procede. Una forma de conseguir que lo sea, es elegirla de forma aleatoria, es decir, al azar aunque en nuestro trabajo diario es habitual obtener los datos, por ejemplo, de los pacientes que ya están en un hospital. En estos casos, podemos admitir que estos pacientes no se han elegido de forma sesgada y que constituyen una muestra representativa de la población en estudio.

Análogamente a lo que pasaba en el capítulo anterior, la *media poblacional* suele representar o caracterizar a una población por lo que es habitual tratar de *estimar* este valor. Si la muestra es representativa de una población, la media aritmética de los datos de esa muestra, a la que denominaremos *media muestral* $\bar{x}$ y que se definirá como la suma de las n observaciones dividido por

el tamaño n de la muestra,

$$\bar{x} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{1}{n} \sum_{i=1}^n X_i$$

será un buen estimador del *parámetro* media poblacional. Este proceso de estimar valores de los parámetros utilizando un estimador (también denominado estadístico) se denomina *Estimación por punto*.

Ejemplo 2.1

Se quiere estimar el tiempo que transcurre desde la administración de la primera dosis de una nueva vacuna contra la hepatitis B hasta que se produce en el individuo una drástica disminución del nivel de anticuerpos contra la mencionada enfermedad, requiriendo éste una nueva dosis de recuerdo.

Para tal fin se eligió una muestra aleatoria de $n = 40$ individuos de la población en estudio en los que se observó el tiempo transcurrido desde la administración de la vacuna hasta la disminución de los anticuerpos, obteniéndose una media muestral $\bar{x} = 35$ días.

En este ejemplo, la población de la que se quieren extraer conclusiones puede ser la población humana y el parámetro de interés puede establecerse en el tiempo medio μ que transcurre desde la administración de la primera dosis de la nueva vacuna en estudio hasta que se produce la drástica disminución del nivel de anticuerpos de la que nos habla el enunciado anterior.

Con objeto de estimar este parámetro, dice el ejemplo que se eligieron al azar 40 individuos a los que se aplicó la vacuna. El tiempo medio muestral de 35 días, se considera una buena estimación del tiempo medio desconocido.

Es fácil entrever en este problema que hay una cierta variación aleatoria en el sentido de que, probablemente, si hubiéramos elegido a otros individuos, la media muestral pudiera haber sido algo distinta o, quizás, muy distinta. Es imprescindible medir esta variabilidad para poder calificar de buenas o malas las conclusiones o estimaciones obtenidas.

La variabilidad aleatoria de los estimadores depende de lo que se esté midiendo. La variabilidad en las medias muestrales de muestras de productos fabricados por una máquina es muy pequeña, puesto que la máquina los hará casi idénticos. En este sentido, la variabilidad de las medias muestrales de estaturas de muestras de individuos dependerá de la variabilidad de estaturas de la población de la que se extraen las muestras: si en la población hay mucha variabilidad, ésta se transmitirá a $\bar{x}$, ocurriendo lo contrario si la población es muy homogénea.

Para formalizar esta cuestión denominemos X a la variable que estemos estudiando, como por ejemplo la estatura de la población en cuestión o, en el ejemplo anterior, el tiempo que transcurre desde la administración de la primera dosis de la vacuna hasta la drástica disminución del nivel de anticuerpos.

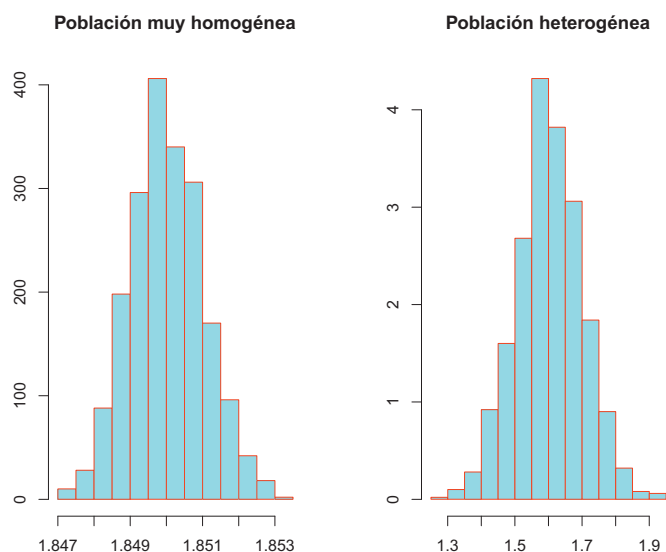


Figura 2.1 : Histogramas de dos poblaciones con distintos grados de concentración

Por centrarnos en el primer caso, pueden ser que casi todos los individuos de la población tengan una estatura muy similar a 1'85 porque la población sea muy homogénea respecto a su estatura, pero puede ser que la población en estudio sea muy rica en cuanto a variedad étnica y que sus estaturas sean muy diversas lo que implicaría mucha dispersión en la población. La variabilidad en la población viene recogida por otro parámetro poblacional que es la *desviación típica poblacional* σ . En el primer caso es probable que el histograma de estaturas de toda la población fuera algo parecido al gráfico de la izquierda de la Figura 2.1 en donde las estaturas están entre 1'84 y 1'86, mientras que en el segundo caso el reparto o distribución de estaturas de la población sea algo similar al histograma de la derecha de la mencionada Figura 2.1 en donde vemos una dispersión de estaturas mayor, al estar éstas entre 1'3 y 2 metros.

2.2. La ley de Probabilidad Normal

En los dos casos mostrados por la Figura 2.1 parece que el histograma tiene una forma acampanada. Este hecho se observó en el siglo XIX y se pensó que le ocurría lo mismo a la mayoría de los fenómenos de la naturaleza por lo que a la ley de probabilidad que se muestra en la Figura 2.2 se la denominó *ley*

de probabilidad normal la cual depende de dos *parámetros*, su media o centro de simetría μ y su desviación típica σ , hablando de la modelización de unos datos por la normal $N(\mu, \sigma)$ lo que representaremos de la forma $X \rightsquigarrow N(\mu, \sigma)$ (por ejemplo una normal de media 10 y desviación típica 2, es decir $X \rightsquigarrow N(10, 2)$) u otros valores de los parámetros. De hecho, con la Estimación por punto o puntual queremos estimar estos dos valores para poder inferir cómo se comporta la población respecto a la característica en estudio.

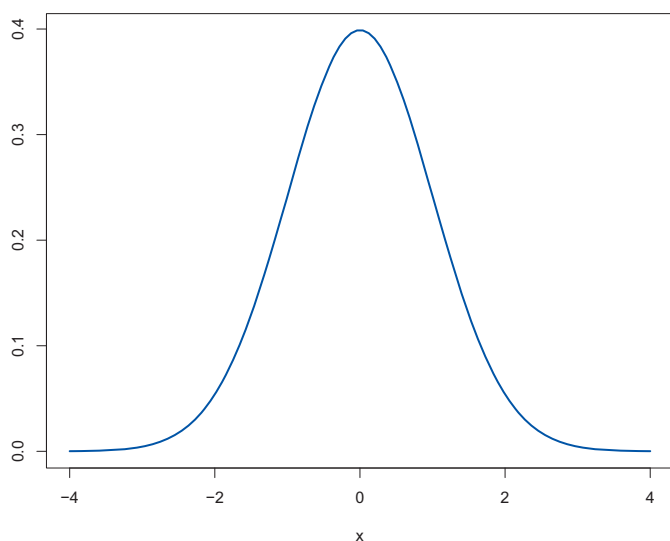


Figura 2.2 : Distribución normal estándar

Si volvemos con el ejemplo de las estaturas, podemos idealizar o, hablando con más propiedad, *modelizar* las dos poblaciones en cuestión por dos leyes normales, la de media 1'85 y desviación típica 0'001 y la distribución normal de media 1'6 y desviación típica 0'1 y sobre impresionarlas en ambos casos, obteniendo la Figura 2.3.

Si fuera correcta esta modelización (y supiéramos Cálculo de Probabilidades) podríamos afirmar por ejemplo que la probabilidad de obtener un individuo mayor de 1'85 en la primera población es 0'5 y que en la segunda es 0'0062. La probabilidad de algo, es decir, de que ocurra un suceso, es un número entre 0 y 1 que indica lo verosímil (valor cercano a 1) o poco verosímil (valor cercano a 0) que es que ocurra ese suceso. Decir que la probabilidad de que llueva mañana es 0'99 nos indica que debemos salir de casa con paraguas porque es *muy probable* que llueva. Si es de 0'01, podemos arriesgarnos a salir de casa sin paraguas.

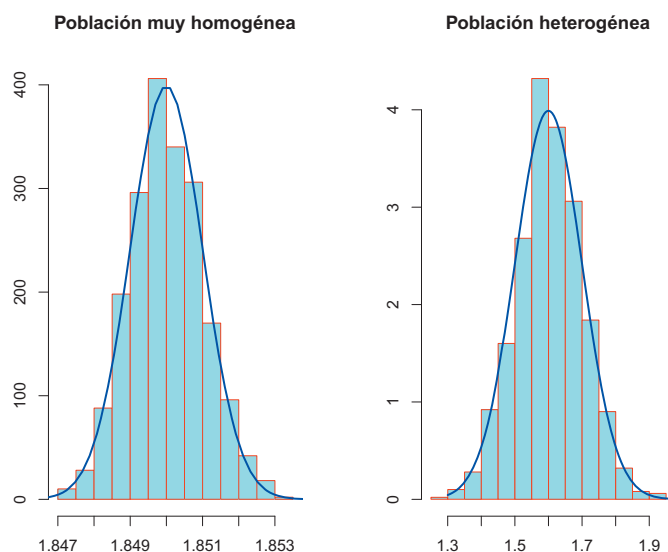


Figura 2.3 : Dos poblaciones con distintos grados de concentración modelizadas con leyes normales

Al hacer estimaciones con la Inferencia Estadística siempre podremos medir la probabilidad de equivocarnos o acertar con dichas inferencias, es decir, podremos valorar nuestras inferencias en términos de probabilidades.

En resumen, cuando analicemos unos datos, lo primero que haremos será modelizar el fenómeno que dio origen a esos datos, puesto que con un estimador transformaremos los datos y la variabilidad o, con más precisión, la *distribución de probabilidad* o *modelo* que rige el fenómeno que dio origen a los datos se transmitirá al estimador que consideremos. Así por ejemplo, si los n datos proceden de una $N(\mu, \sigma)$, la distribución o modelo que rige a la media muestral $\bar{x}$ es una $N(\mu, \sigma/\sqrt{n})$ lo que permite (al igual que antes) calcular probabilidades de obtener valores mayores o menores que un valor determinado o, simplemente, ver que a medida que aumentamos en tamaño n de la muestra, la distribución de la media muestral está más concentrada alrededor de la media puesto que la desviación típica viene dividida por dicho valor.

Ejemplo 2.1 (continuación)

Por datos recogidos de experimentos similares con otras vacunas, se modelizó a la variable $X =$ tiempo que transcurre desde la administración de la primera dosis de la vacuna hasta la drástica disminución del nivel de anticuerpos, mediante una distribución normal de media 33 días y desviación típica 7 días, es decir, una $N(33, 7)$.

Como la Inferencia Estadística determinó que siempre que tengamos una variable X con distribución $N(\mu, \sigma)$ la media muestral de datos extraídos de dicha población sigue una ley

$N(\mu, \sigma/\sqrt{n})$, en estudio de esta vacuna podemos decir que la media muestral $\bar{x}$ sigue una $N(33, 7/\sqrt{40}) = N(33, 1'1068)$.

Aunque hoy en día ya sabemos que la ley de probabilidad normal rige los fenómenos de la naturaleza tan habitualmente como otras distribuciones, dado que gran parte de la Inferencia Estadística se construyó en los siglos pasados admitiendo este modelo, va a ser necesario conocerlo más a fondo y saber calcular probabilidades relacionadas con él. A esto dedicaremos la siguiente sección.

La distribución Normal fue propuesta por primera vez como modelo probabilístico por De Moivre en 1733 y por Laplace, de forma independiente, en 1774 pero la referencia más utilizada en relación con la distribución que nos ocupa es la de Laplace (1814) y Gauss (1809) en donde la utilizaron en el análisis de los errores en Astronomía y Geodesia aunque el nombre de *normal* se debe a Quetelet.

Ya hemos visto su forma general en la Figura 2.2. Variando su dos parámetros, media μ y desviación típica σ , la deslizaremos por el eje de abscisas y la haremos más o menos puntiaguda pues la masa de probabilidad que hay bajo esa curva es siempre igual a 1. Pero hay unos valores de μ y σ muy importantes, son $\mu = 0$ y $\sigma = 1$ que hacen que la normal $N(0, 1)$ se denomine de una forma particular: *normal estándar*. Cualquier distribución normal, y por tanto cualquier grupo de datos procedentes de una normal no estándar, pueden *tipificarse* o *estandarizarse* es decir, convertirse en datos procedentes de una normal estándar, restándoles su media y dividiéndolos por su desviación típica.

Matemáticamente esto significa que si X es una variable con distribución modelo $N(\mu, \sigma)$, la variable

$$Z = \frac{X - \mu}{\sigma}$$

sigue una distribución normal estándar $N(0, 1)$.

En el cálculo de probabilidades bajo la curva normal es muy frecuente querer calcular probabilidades hasta un determinado punto, como el área roja de la Figura 2.4 es decir, el área acumulada hasta, en este caso, la abscisa $x = -0'7$. Aunque hasta hace muy poco tiempo estas probabilidades se calculaban mediante una tablas de probabilidades, hoy en día es más sencillo y preciso calcularlas con R, ejecutando en este caso, dado que es un modelo $N(0, 1)$ el de la figura,

```
> pnorm(-0.7, 0, 1)
[1] 0.2419637
```

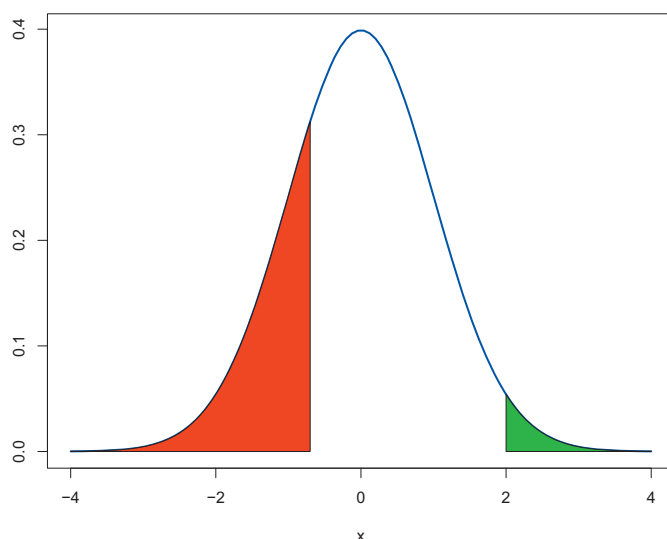


Figura 2.4 : Áreas bajo la curva Normal

lo que indica que el área de probabilidad acumulada hasta $x = -0.7$ es 0.2419637.

También es muy frecuente calcular *probabilidades cola*, es decir, áreas bajo la normal después de un determinado valor, como la zona verde de la Figura 2.4, en este caso, a la derecha de la abscisa $x = 2$. Dado que, como dijimos antes, el área acumulada bajo la curva normal es igual a 1, el valor buscado se calculará ejecutando

```
> 1-pnorm(2,0,1)
[1] 0.02275013
```

Es decir, la probabilidad cola más allá de la abscisa $x = 2$ es 0.02275.

Otro cálculo muy habitual relacionado con la normal es lo que se denomina *cuantil*, que es el inverso de las probabilidades antes calculadas. Es decir, en lugar de calcular la probabilidad acumulada hasta un determinado valor, se quiere determinar el valor de la abscisa que tiene acumulada una determinada probabilidad α hasta él, valor que se denomina α -cuantil. Por ejemplo, por los cálculos anteriores, la abscisa -0.7 es el 0.24196-cuantil aunque los α -cuantiles más buscados son el 0.05-cuantil y el 0.95-cuantil. Con R todos los cuantiles

son muy fáciles de determinar con la función `qnorm`. Por ejemplo, el 0'24196-cuantil de la $N(0, 1)$ se determina ejecutando

```
> qnorm(0.24196, 0, 1)
[1] -0.7000117
```

Si la distribución normal considerada no fuera la $N(0, 1)$ sino otra normal con otros parámetros, en todos los cálculos anteriores bastaría cambiar el 0 y el 1 del segundo y tercer argumento para hacer los correspondientes cálculos para ese modelo. De hecho, cuando se ejecutan cálculos con una $N(0, 1)$ no es necesario poner estos valores, R los toma por defecto. Por ejemplo, el 0'95-cuantil de una $N(1, 2)$ sería

```
> qnorm(0.95, 1, 2)
[1] 4.289707
```

Es decir, que 4'2897 es el valor de la abscisa de una $N(1, 2)$ que deja a la izquierda un área de probabilidad 0'95 o, equivalentemente pues el área bajo toda curva normal es igual a 1, es el valor que deja a su derecha un área de probabilidad 0'05.

En los libros de Estadística, suele denotarse por z_α al valor de la abscisa de una $N(0, 1)$ que deja a la derecha una probabilidad α y, lógicamente, $z_{\alpha/2}$ al valor de la abscisa de una $N(0, 1)$ que deja a la derecha una probabilidad $\alpha/2$.

Ejemplo 2.1 (continuación)

Dado que hemos modelizado nuestros datos por una $N(33, 7)$, lo que implica que para una muestra de 40 datos la media muestral se distribuya como una $N(33, 1'1068)$, nos podemos preguntar por lo probable que resulta obtener una media muestral de 35 días o mayor. Matemáticamente lo expresaríamos como

$$P\{\bar{x} > 35\}$$

o, tipificando, es decir, restando la media y dividiendo por la desviación típica en ambos lados de la desigualdad para que los dos sucesos tengan la misma probabilidad,

$$P\{\bar{x} > 35\} = P\left\{\frac{\bar{x} - 33}{1'1068} > \frac{35 - 33}{1'1068}\right\} = P\{Z > 1'807\}$$

en donde Z es una variable con distribución normal estándar es decir, $N(0, 1)$. Ambas probabilidades, que deben de ser iguales, se calculan fácilmente con R,

```
> 1-pnorm(35, 33, 1.1068)
[1] 0.03538
```

```
> 1-pnorm(1.807)
[1] 0.03538
```

Con objeto de practicar más en el cálculo de probabilidades y cuantiles relacionados con una distribución normal, incluimos el siguiente ejemplo en el que recomendamos al lector que haga un dibujo semejante a la Figura 2.4, sombreando las áreas de probabilidad que va a calcular o marcando la abscisa que va a determinar.

Ejemplo 2.2

Si Z es una variable que sigue una distribución $N(0, 1)$, obtenemos los siguientes valores:
 $P\{Z < 2'03\} = 0'9788$, ya que

```
> pnorm(2.03)
[1] 0.9788217
```

$P\{Z < -0'3\} = 0'3821$, ya que

```
> pnorm(-0.3)
[1] 0.3820886
```

$P\{Z > -1'39\} = 0'9177$, ya que

```
> 1-pnorm(-1.39)
[1] 0.9177356
```

$P\{-1'2 < Z < 1'05\} = P\{Z < 1'05\} - P\{Z < -1'2\} = 0'738$, ya que

```
> pnorm(1.05)-pnorm(-1.2)
[1] 0.7380713
```

$P\{1'68 < Z < 3'36\} = P\{Z < 3'36\} - P\{Z < 1'68\} = 0'0461$, ya que

```
> pnorm(3.36)-pnorm(1.68)
[1] 0.04608895
```

$P\{-1'2 < Z < -0'03\} = P\{0'03 < Z < 1'2\} = 0'3729$, ya que

```
> pnorm(-0.03)-pnorm(-1.2)
[1] 0.3729639
```

Si X sigue una $N(3, 2)$, las probabilidades correspondientes a esta distribución se pueden determinar primero tipificando y después por la búsqueda de la probabilidad tipificada o directamente. Así por ejemplo,

$$P\{X < 1'5\} = P\{Z < (1'5 - 3)/2\} = P\{Z < -0'75\} = 0'2266$$

ya que

```
> pnorm(1.5,3,2)
[1] 0.2266274

> pnorm((1.5-3)/2)
[1] 0.2266274
```

Por último, si queremos conocer el z tal que $P\{Z > z\} = 0'01$, es decir, el 0'99-cuantil de la normal estándar, debemos ejecutar

```
> qnorm(0.99)
[1] 2.326348
```

2.3. La distribución t de Student

En el Ejemplo 2.2 suponíamos que la variable en estudio X seguía una distribución $N(33, 7)$, pero es poco verosímil admitir que conocemos la desviación típica poblacional σ y, si no la conocemos, la distribución de la media muestral $\bar{x}$, cuya desviación típica es $\sigma/\sqrt{n}$, dependerá del parámetro desconocido σ y no podrá ser utilizada.

Si en lugar de la distribución estandarizada de $\bar{x}$

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

que será una $N(0, 1)$ si los datos proceden de una $N(\mu, \sigma)$, sustituimos σ por la cuasidesviación típica muestral S , la distribución de

$$\frac{\bar{x} - \mu}{S/\sqrt{n}}$$

fue estudiada y tabulada por W.S. Gosset que la publicó en 1908 bajo el pseudónimo de Student por lo que se conoce bajo el nombre de *distribución t de Student*.

Esta distribución sólo depende del denominado número de *grados de libertad* que es $n - 1$ en el caso de más arriba que estemos estudiando la distribución de la media muestral de n datos por lo que se habla en este caso de una t_{n-1} .

Su forma es muy similar a la normal. En la Figura 2.5 aparece un distribución modelo t de Student con 12 grados de libertad, es decir, una t_{12} .

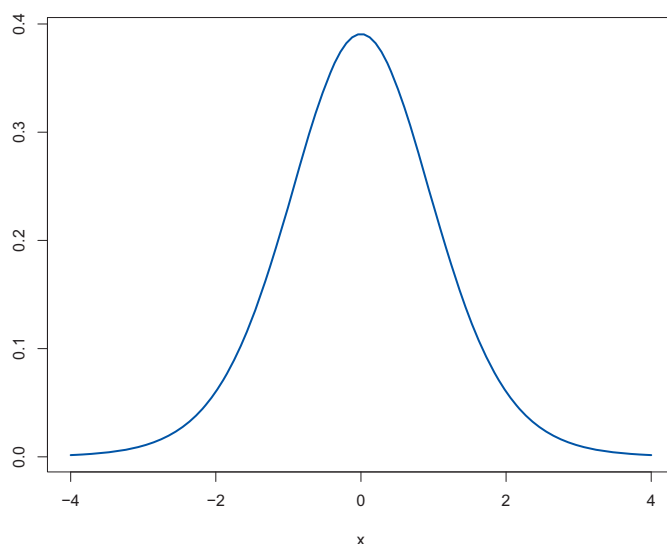


Figura 2.5 : Distribución t de Student

El cálculo de áreas bajo una t de Student y de cuantiles asociados a esta distribución se hace con R muy fácilmente de forma análoga al caso de la normal pero ahora utilizando, respectivamente, las funciones de R $\text{pt}(x, n)$ en el caso de probabilidades acumuladas hasta el punto x por una t de Student con n grados de libertad y por la función $\text{qt}(p, n)$ en el caso de que queramos determinar el cuantil de una t de Student con n grados de libertad que acumula un área p bajo dicha curva.

Matemáticamente, el valor de una abscisa de una t_n de Student que deja a la derecha un área α se denomina $t_{n;\alpha}$.

Ejemplo 2.3

El área acumulada hasta la abscisa $x = 1,3$ por una distribución t_{10} de Student es 0'88861 ya que

```
> pt(1.3,10)
[1] 0.8886171
```

y el área que deja a la derecha de $x = 1,1$ una distribución t_5 de Student será 0'1607 ya que

```
> 1-pt(1.1,5)
[1] 0.1607254
```

Por último, el valor de una abscisa de una distribución t_{11} de Student con 11 grados de libertad que deja a su derecha un área igual a 0'025 será $t_{11;0'025} = 2'201$ ya que

```
> qt(0.975,11)
[1] 2.200985
```

Tanto se parece la t de Student a una normal que, cuando el número de grados de libertad es mayor que 30 apenas si se diferencian como puede verse en la Figura 2.6.

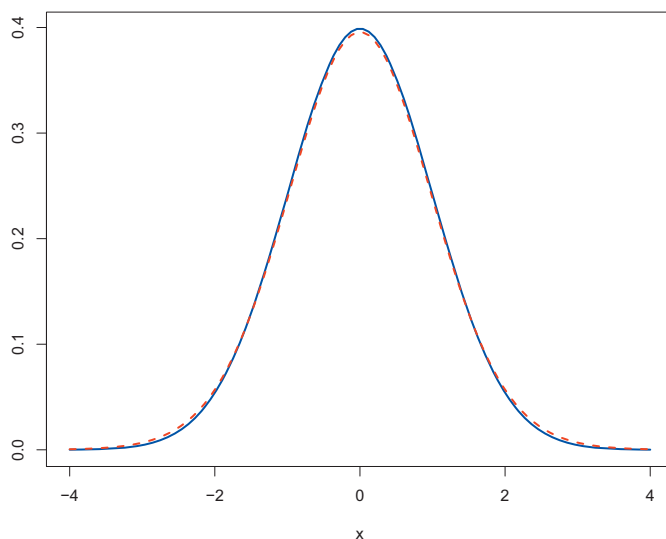


Figura 2.6 : Distribución normal estándar $N(0,1)$ en azul y t_{30} de Student en rojo

Ejemplo 2.4

Por ejemplo comparemos el área acumulada hasta la abscisa $x = 1$ por ambas distribuciones,

```
> pnorm(1)
[1] 0.8413447
> pt(1,30)
[1] 0.8373457
```

Apenas si hay diferencias, las cuales disminuirán a medida que aumenten los grados de libertad.

Esto tendrá interesantes aplicaciones en la estimación de la media poblacional.

2.4. Estimación de la media poblacional

Resumiendo lo estudiado en las secciones anteriores, si los datos proceden de una distribución $N(\mu, \sigma)$, el estimador que debemos utilizar en la estimación de la media poblacional μ es la media muestral $\bar{x}$, estadístico que tendrá una distribución $N(\mu, \sigma/\sqrt{n})$, es decir, tipificando

$$\frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

será una $N(0, 1)$. Pero si la desviación típica de la población es desconocida, $\bar{x}$ tendrá una distribución t_{n-1} . Más en concreto,

$$\frac{\bar{x} - \mu}{S/\sqrt{n}}$$

tendrá una distribución t_{n-1} .

Y todo esto si los tamaños muestrales son pequeños, porque si n es grande, bien por el comportamiento límite de la distribución t de Student o por lo que se denomina *Teorema Central del Límite*, aunque los datos no procedan una distribución normal, se puede utilizar que

$$\frac{\bar{x} - \mu}{S/\sqrt{n}}$$

sigue aproximadamente una distribución $N(0, 1)$.

Ejemplo 2.5

Se supone que la longitud craneal de los individuos de una población sigue una distribución normal con una desviación típica de 12'7 mm. Si elegimos de esa población al azar 10

individuos, la probabilidad de que la media de esa muestra difiera de la poblacional en más de 4'4 mm. será

$$P\{|\bar{x} - \mu| > 4'4\} = P\{|Z| > 1'1\} = 2 \cdot 0'1357 = 0'2714$$

por ser

$$\frac{\bar{x} - \mu}{12'7/\sqrt{10}} \rightsquigarrow N(0, 1)$$

y

```
> 1-pnorm(1.1)
[1] 0.1356661
```

Si hubiera sido desconocida la varianza poblacional y la muestra nos hubiera dado una cuasidesviación típica $S = 12$, la probabilidad buscada sería,

$$P\{|\bar{x} - \mu| > 4'4\} = P\{|t_9| > 1'1595\} = 2 \cdot P\{t_9 > 1'1595\} = 2 \cdot 0'1380 = 0'276$$

al tener que utilizar una t de Student, por ser la varianza poblacional desconocida y las muestras pequeñas,

$$\frac{\bar{x} - \mu}{S/\sqrt{10}} \rightsquigarrow t_9$$

y ser

```
> 1-pt(1.1595,9)
[1] 0.1380443
```

Ejemplo 2.6

Con objeto de estimar los niveles de hierro en la sangre de los varones adultos sanos, se obtuvo una muestra de tamaño 100 que proporcionó una cuasidesviación típica de 15 microgramos por cada 100ml de sangre. La probabilidad de que la media de esa misma muestra difiera de la media poblacional en más de 3 microg/100ml será

$$P\{|\bar{x} - \mu| > 3\} = P\{|Z| > 2\} = 0'0455$$

por ser

```
> 2*(1-pnorm(2))
[1] 0.04550026
```

2.5. Estimación de la varianza poblacional: Distribución χ^2 de Pearson

Al igual que la media de la muestra es un buen estimador de la media de la población, la cuasivarianza muestral S^2 definida en el capítulo anterior es un buen estimador del parámetro varianza poblacional σ^2 , por lo que su raíz cuadrada, la cuasidesviación típica muestral S es un buen estimador de la desviación típica poblacional σ .

De nuevo, para hacer inferencias en base a este estimador necesitamos conocer su distribución surgiendo así la denominada *distribución χ^2 de Pearson* que, al igual que la distribución t de Student también depende de un parámetro denominado *grados de libertad*, siendo esta distribución asimétrica aunque siempre tomando valores positivos. Su forma es la dada por la Figura 2.7.

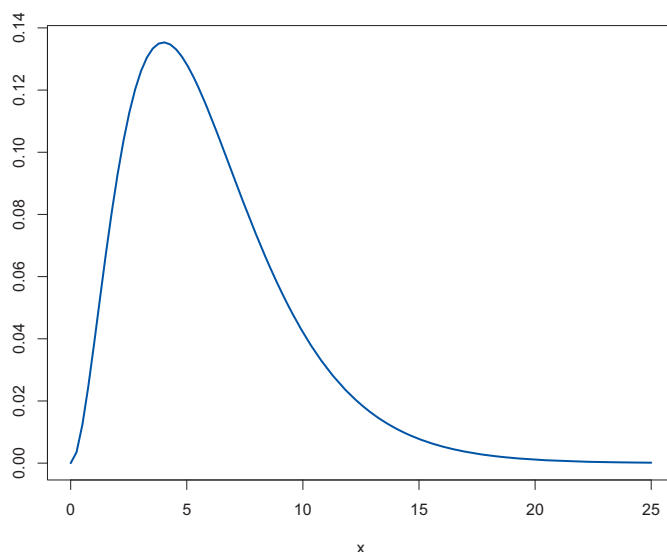


Figura 2.7 : Distribución χ^2 de Pearson

Pues bien, si los n datos observados $X_1, \dots, X_n$ proceden de una $N(\mu, \sigma)$, entonces, la distribución de cuasivarianza muestral S^2 , estandarizada, es decir,

$$\frac{(n-1)S^2}{\sigma^2}$$

es una distribución χ^2 de Pearson con $n-1$ grados de libertad, es decir una χ_{n-1}^2 .

Las probabilidades acumuladas hasta un punto x por una χ_n^2 se calculan

con R mediante la función `pchisq(x,n)` y los α -cuantiles, es decir, el valor de una abscisa de una χ_n^2 que deja a la derecha un área de probabilidad α se representa matemáticamente por $\chi_{n;\alpha}^2$, se calcula con la función de R `qchisq(1 -  $\alpha$ , n)`.

Ejemplo 2.7

Calcular la probabilidad de que en un recuento de glóbulos blancos en individuos de una muestra aleatoria simple de tamaño 10, la cuasivarianza muestral sobrestime a la varianza poblacional en más de un tercio de su valor, suponiendo que el número de glóbulos blancos sigue una distribución normal.

La probabilidad pedida será, después de multiplicar por $n - 1 = 9$ y dividir ambos miembros de la desigualdad por σ ,

$$P\{S^2 > \sigma^2 + \sigma^2/3\} = P\{9 \cdot S^2 \sigma^2 > 9\sigma^2(1 + 1/3)/\sigma^2\} = \chi_9^2 > 12\} = 0.2133$$

ya que

```
> 1-pchisq(12,9)
[1] 0.2133093
```

2.6. Estimación del cociente de varianzas poblacionales: Distribución F de Snedecor

Cuando comparemos dos grupos de datos procedentes de dos poblaciones con distribuciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, resultará necesario analizar si puede admitirse que las varianzas de ambas poblaciones pueden considerarse iguales o, equivalentemente, si puede admitirse que su cociente σ_1^2/σ_2^2 es igual a 1.

Este cociente de varianzas poblacionales se estimará con el cociente de cuasivarianzas muestrales S_1^2/S_2^2 procedentes de dos muestras de tamaños n_1 y n_2 de cada una de las dos poblaciones en estudio. Pues bien, el cociente

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

sigue una distribución conocida como *distribución F de Snedecor* con $(n_1 - 1, n_2 - 1)$ grados de libertad. Su forma es la de la Figura 2.8, muy parecida a una distribución χ^2 . De hecho, una distribución F de Snedecor con (n_1, n_2) grados de libertad, distribución que se representa por $F_{(n_1, n_2)}$ se puede obtener como el cociente de dos distribuciones χ^2 independientes con grados de libertad n_1 la del numerador y n_2 la del denominador.

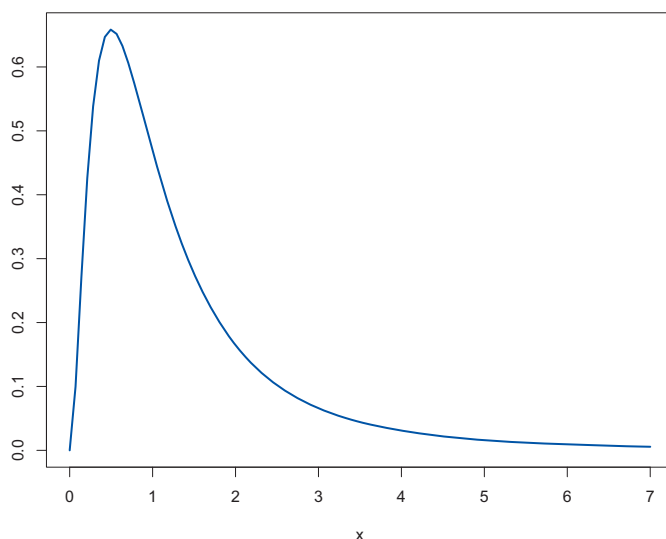


Figura 2.8 : Distribución F de Snedecor

Las probabilidades acumuladas hasta un punto x por una $F_{(n_1, n_2)}$ se calculan con R mediante la función `pf(x, n1, n2)` y los α -cuantiles, es decir, el valor de una abscisa de una $F_{(n_1, n_2)}$ que deja a la derecha un área de probabilidad α se representa matemáticamente por $F_{n_1, n_2; \alpha}$, se calcula con la función de R, `qf(1 - alpha; n1, n2)`.

Ejemplo 2.8

Un investigador supone que los niveles de vitamina A en dos poblaciones humanas independientes se distribuyen normalmente con el mismo nivel medio y varianzas iguales $\sigma_1^2 = \sigma_2^2$. Extraída una muestra aleatoria de cada población de tamaños $n_1 = 10$ y $n_2 = 12$ respectivamente, se obtuvieron como cuasivarianzas muestrales los valores $S_1^2 = 955$ y $S_2^2 = 415'2$. ¿Qué probabilidad habría de haber observado un desequilibrio entre las cuasivarianzas muestrales mayor del obtenido $955/415'2 = 2'3$?

Como las varianzas poblacionales se suponen iguales es decir, suponemos que es $\sigma_1^2 = \sigma_2^2$, será

$$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = S_1^2/S_2^2$$

y seguirá este cociente una distribución $F_{(9, 11)}$. La probabilidad pedida será,

$$P\left\{\frac{S_1^2}{S_2^2} > 2'3\right\} = P\{F_{(9, 11)} > 2'3\} = 0'09696$$

ya que

```
> 1-pf(2.3,9,11)
[1] 0.09695708
```

Capítulo 3

Estimación por Intervalos de Confianza

3.1. Introducción

En el capítulo anterior estudiamos la *Estimación por punto* de las características o *parámetros* de la población que queremos investigar y así dijimos que, si queremos estimar la media μ de una población, debemos utilizar la media $\bar{x}$ de una muestra representativa extraída de la población en estudio. No obstante, raramente la estimación por punto coincidirá exactamente con el parámetro a estimar, es decir, rara vez la media de la muestra seleccionada al azar será tal que $\bar{x} = \mu$. Sin duda, es mucho más interesante realizar la inferencia con un intervalo de posibles valores del parámetro —al que denominaremos *Intervalo de Confianza*—, de manera que, antes de tomar la muestra, el desconocido valor del parámetro se encuentre en dicho intervalo con una probabilidad todo lo alta que deseemos.

Así por ejemplo, es mucho más deseable afirmar que la media poblacional μ está entre $\bar{x} - 0'1$ y $\bar{x} + 0'1$, con probabilidad 0'99, que dando un valor concreto como estimación puntual de μ , el cual es posible que esté muy alejado del verdadero.

Con objeto de aumentar la precisión de la inferencia, serán deseables intervalos de confianza lo más cortos posible.

No obstante, la longitud del intervalo de confianza dependerá de lo alta que queramos sea la probabilidad con la que dicho intervalo —cuyos extremos son aleatorios— cubra a μ y, por tanto, del modelo que elijamos para explicar la variable en estudio. Así por ejemplo si queremos determinar el intervalo de confianza para la media de una población normal de varianza conocida σ , éste será

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

en donde $z_{\alpha/2}$ es, como dijimos en el capítulo anterior, el valor de la abscisa de una $N(0, 1)$ que deja a su derecha —bajo la función de densidad— un área de probabilidad $\alpha/2$.

Como se ve, la longitud del intervalo de confianza, es decir, la diferencia entre el extremo superior y el inferior,

$$2 \cdot z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

depende de la probabilidad $1 - \alpha$ elegida en su construcción, a la que denominaremos *coeficiente de confianza*, y del tamaño muestral (a mayor tamaño muestral n , menor será la longitud del intervalo).

Para un tamaño muestral fijo, cuanto mayor sea el coeficiente de confianza, más grande será $z_{\alpha/2}$ y por tanto, mayor su longitud. Por tanto, antes de construir un intervalo de confianza, habrá que prefijar cuidadosamente el valor del coeficiente de confianza de manera que la *probabilidad con la que confiamos* el intervalo cubra al desconocido valor del parámetro sea alta, pero conservando inferencias válidas.

Así, de poco interés resultará concluir que hay probabilidad 0'999 de que el intervalo (en metros) $[\bar{x} - 2, \bar{x} + 2]$, cubra la estatura media de la población.

Los coeficientes de confianza que se suelen considerar son 0'90, 0'95 y 0'99, aunque esto dependerá del investigador, el cual deberá tener siempre en cuenta los comentarios anteriores. Por ejemplo, una varianza poblacional σ^2 pequeña o un tamaño muestral grande pueden permitir un mayor coeficiente de confianza sin un aumento excesivo de la longitud del intervalo.

Formalmente definimos el intervalo de confianza para un parámetro θ de la siguiente manera.

Definición

Supongamos que X es la variable aleatoria en estudio, cuya distribución depende de un parámetro desconocido θ , y $X_1, \dots, X_n$ una muestra aleatoria simple de dicha variable.

Si $T_1(X_1, \dots, X_n)$ y $T_2(X_1, \dots, X_n)$ son dos estadísticos tales que

$$P\{T_1(X_1, \dots, X_n) \leq \theta \leq T_2(X_1, \dots, X_n)\} = 1 - \alpha$$

el intervalo

$$[T_1(x_1, \dots, x_n), T_2(x_1, \dots, x_n)]$$

recibe el nombre de *Intervalo de Confianza* para θ de *coeficiente de confianza* $1 - \alpha$.

Obsérvese que tiene sentido hablar de que, antes de tomar la muestra, el intervalo aleatorio

$$[T_1(X_1, \dots, X_n) , T_2(X_1, \dots, X_n)]$$

cubra al verdadero y desconocido valor del parámetro θ con probabilidad $1 - \alpha$ pero, una vez elegida una muestra particular $x_1, \dots, x_n$, el intervalo no aleatorio

$$[T_1(x_1, \dots, x_n) , T_2(x_1, \dots, x_n)]$$

cubrirá o no a θ , pero ya no tiene sentido hablar de la probabilidad con que lo cubre.

Es decir, podemos hacer afirmaciones del tipo de que en un $100(1 - \alpha) \%$ de las veces, el intervalo que obtengamos cubrirá al parámetro, pero nunca de que, por ejemplo, hay probabilidad $1 - \alpha$ de que el intervalo de confianza $[1'65, 1'83]$ cubra al parámetro, ya que los extremos de este último intervalo —y como siempre el parámetro— son números y no variables aleatorias.

Obsérvese también que el intervalo de confianza es un subconjunto de los posibles valores del parámetro precisamente por ser no aleatorio.

Así mismo mencionemos que cualquier par de estimadores T_1 y T_2 que cumplan la condición impuesta en la definición anterior darán lugar a un intervalo de confianza. Habitualmente éstos serán dos funciones del estimador natural obtenido para cada caso en el capítulo anterior. De hecho, en las siguientes secciones indicaremos cuál es el intervalo de confianza que razonablemente debe utilizarse en cada situación concreta. En muchos casos su obtención se hará utilizando un paquete estadístico y, en otras, aplicando las fórmulas que se indica por lo que incluiremos ejemplos de ambas situaciones.

Recordamos la notación que utilizaremos, tanto en los intervalos de confianza como en el resto del libro: denotaremos por z_p , $t_{n;p}$, $\chi^2_{n;p}$ y $F_{n_1, n_2; p}$, respectivamente, el valor de la abscisa de una distribución $N(0, 1)$, t_n de Student, χ^2_n de Pearson y F_{n_1, n_2} de Snedecor, que deja a su derecha —bajo la correspondiente función de densidad— un área de probabilidad p .

3.1.1. Cálculo de Intervalos de Confianza con R

En el capítulo siguiente veremos que el intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test bilateral. Por esta razón se utiliza la misma función de R para obtener intervalos de confianza y test de hipótesis sobre un parámetro.

En concreto, la función de R que nos va a proporcionar los intervalos (y los tests), es la función `t.test`. Con ella vamos a poder determinar los Intervalos

de Confianza (y tests) para la media, para datos apareados y para la diferencia de medias, pero no para aquellos casos en los que la varianza, varianzas o medias poblacionales sean conocidas sino para cuando haya que estimarlas a partir de los datos. También queremos advertir que, para poder aplicar esta función, es necesario conocer los datos individualmente ya que no podremos utilizarla cuando sólo conozcamos los valores de las medias o cuasivarianzas muestrales y no los datos de donde éstas proceden.

La función a utilizar en el caso de Intervalos de Confianza es

```
t.test(x, y = NULL, paired = FALSE, var.equal = FALSE, conf.level = 0.95)
```

Entrando a describir cada uno de sus argumentos, en primer lugar diremos que los valores que aparecen después del símbolo = son los que toma la función por defecto y que, por tanto, no será necesario especificar si son los valores que deseamos ejecutar. En `x` incorporamos los datos de la muestra, si se trata de inferencias para una sola muestra; si se trata de datos apareados o de dos muestras independientes, introduciremos los datos de la segunda muestra en el argumento `y`.

Si especificamos `paired=F` (lo cual no es necesario puesto que es la opción tomada por defecto), estamos es una situación de datos no apareados. Un caso de datos apareados debe especificarse con `paired=T`.

El argumento `var.equal` nos permite indicar qué tipo de situación tenemos en el caso de comparación de dos poblaciones independientes. Si es `var.equal=T` tendremos una situación en la que las varianzas de ambas poblaciones se suponen iguales, y el intervalo será el habitual basado en una t de Student. Si especificamos `var.equal=F` las varianzas de ambas poblaciones no se suponen iguales y, en ese caso, estamos requiriendo un intervalo basado en una t de Student pero en donde los grados de libertad se determina por la aproximación de Welch.

El último argumento permite especificar el coeficiente de confianza, tomándose por defecto el valor 0'95.

El intervalo de confianza para el cociente de varianzas poblacionales se obtiene con la función

```
var.test(x, y, conf.level = 0.95)
```

en donde incorporamos los datos en los argumentos `x` e `y`. De nuevo aquí necesitaremos conocer los datos concretos y no admite esta función la situación de ser las medias poblacionales conocidas.

3.2. Intervalo de confianza para la media de una población normal

Tanto en esta sección como en las siguientes, determinaremos intervalos de confianza de *colas iguales*. Es decir, aquellos tales que, si el coeficiente de confianza es $1 - \alpha$, dejan en cada uno de los extremos la mitad de la probabilidad, $\alpha/2$.

En esta sección suponemos que los n datos proceden de una población $N(\mu, \sigma)$, y lo que pretendemos determinar es el intervalo de confianza para la media μ .

Como vimos en la Sección 2.4, en esta situación, tanto si la varianza poblacional σ^2 es conocida como si no lo es, el estimador natural de μ es la media muestral $\bar{x}$.

σ conocida

El intervalo buscado será

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right].$$

σ desconocida

En este caso de que la varianza poblacional sea desconocida, el intervalo de confianza para la media resulta

$$\left[\bar{x} - t_{n-1; \alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1; \alpha/2} \frac{S}{\sqrt{n}} \right]$$

en donde S^2 es la cuasivarianza muestral.

Ejemplo 3.1

Un terapeuta desea estimar, con una confianza del 99 %, la fuerza media de un músculo determinado en los individuos de una población. Admitiendo que las unidades de fuerza siguen una distribución normal de varianza 144, seleccionó una muestra aleatoria de 25 individuos de la población, para la que obtuvo una media muestral de $\bar{x} = 85$.

Como no tenemos los datos observados, en este caso deberemos utilizar las fórmulas anteriores para calcular el intervalo de confianza. En estas condiciones, el intervalo de confianza será

$$\left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right] = \left[85 - z_{0'01/2} \frac{12}{\sqrt{25}}, 85 + z_{0'01/2} \frac{12}{\sqrt{25}} \right]$$

Como es $z_{0'01/2} = z_{0'005}$ es valor de una abscisa de una normal estándar $N(0, 1)$ que deja a la derecha un área de probabilidad 0'005, este valor se calculará, como vimos en la Sección 2.2, ejecutando

```
> qnorm(1-0.005)
[1] 2.575829
```

Por tanto, el intervalo de confianza buscado será,

$$\left[85 - 2'575829 \frac{12}{\sqrt{25}}, 85 + 2'575829 \frac{12}{\sqrt{25}} \right] = [78'82, 91'18].$$

Estos cálculos los puede obtener con una calculadora o con R ejecutando

```
> 85-2.575829*12/sqrt(25)
[1] 78.81801
```

```
> 85+2.575829*12/sqrt(25)
[1] 91.18199
```

Si, como es más razonable, el terapeuta no supone conocida la varianza poblacional, deberá estimarla con la cuasivarianza muestral de los 25 individuos seleccionados. Si ésta fue $S^2 = 139$, el intervalo de confianza será

$$\left[85 - t_{24;0'01/2} \sqrt{\frac{139}{25}}, 85 + t_{24;0'01/2} \sqrt{\frac{139}{25}} \right] = [78'4, 91'59]$$

ya que el valor de la abscisa de una t de Student con 24 grados de libertad que deja a la derecha un área de probabilidad $0'01/2 = 0'005$ será (vea la Sección 2.3),

```
> qt(1-0.005,24)
[1] 2.79694
```

y es

```
> 85-2.79694*sqrt(139/25)
[1] 78.40491
```

```
> 85+2.79694*sqrt(139/25)
[1] 91.59509
```

Ejemplo 3.2

Una muestra aleatoria de 10 clientes de una farmacia determinada mostró los siguientes tiempos de espera hasta que son atendidos, en minutos:

2, 10, 4, 5, 1, 0, 5, 9, 3, 9

Determinar un intervalo de confianza, con coeficiente de confianza 0'9, para el tiempo medio de espera, admitiendo que el tiempo de espera en esa farmacia sigue una distribución normal. Se trata de calcular el intervalo de confianza para la media de una población normal de varianza desconocida que vimos era

$$\left[\bar{x} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right].$$

De los datos del enunciado se desprende que es $\bar{x} = 4'8$ y $S = 3'52$, como fácilmente se obtiene con R,

```
> x<-c(2,10,4,5,1,0,5,9,3,9)
> mean(x)
[1] 4.8
> sd(x)
[1] 3.521363
```

Por tanto, como además es $t_{n-1;\alpha/2} = t_{9;0'05} = 1'833$ ejecutando

```
> qt(1-0.05,9)
[1] 1.833113
```

el intervalo de confianza solicitado será

$$\left[\bar{x} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right] = \left[4'8 - 1'833 \frac{3'52}{\sqrt{10}}, 4'8 + 1'833 \frac{3'52}{\sqrt{10}} \right] = [2'76, 6'84].$$

Si queremos obtener el intervalo directamente con R, ejecutaríamos

```
> t.test(x,conf.level=0.9)
One Sample t-test
```

```
data: x
t = 4.3105, df = 9, p-value = 0.00196
alternative hypothesis: true mean is not equal to 0
90 percent confidence interval:
 2.758732 6.841268
sample estimates:
mean of x
 4.8
```

(1)

obteniendo en (1) el mismo intervalo que antes.

3.3. Intervalo de confianza para la media de una población no necesariamente normal. Muestras grandes

Si el tamaño de la muestra es lo suficientemente grande (digamos mayor que 30 datos), el intervalo de confianza se basará siempre en una normal, sea

o no conocida la varianza de la población y procedan o no los datos de una normal. En concreto,

Si σ es conocida el intervalo de confianza para μ de coeficiente de confianza $1 - \alpha$ será

$$I = \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

y si σ es desconocida

$$I = \left[\bar{x} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

siendo, como antes, S la cuasidesviación típica muestral.

Ejemplo 3.3

Los siguientes datos son valores de actividad (en micromoles por minuto por gramo de tejido) de una cierta enzima observada en el tejido gástrico de 35 pacientes con carcinoma gástrico

0'360	1'185	0'524	0'870	0'356	2'567	0'566
1'789	0'578	0'578	0'892	0'345	0'256	0'987
0'355	0'989	0'412	0'453	1'987	0'544	0'798
0'634	0'355	0'455	0'445	0'755	0'423	0'754
0'452	0'452	0'450	0'511	1'234	0'543	1'501

El histograma de estos datos (Figura 3.1) muestra claramente una fuerte asimetría a la derecha, lo cual sugiere que los valores de actividad no siguen una distribución normal.

No obstante, al ser el tamaño muestral bastante grande la media muestral $\bar{x}$ sí sigue una distribución normal. Es decir, si hiciéramos un histograma en el que representáramos los valores obtenidos por la media muestral en un gran número de muestras, éste tendría forma acampanada aunque, como ocurre en este caso, la variable poblacional no siga una distribución normal.

El intervalo de confianza a utilizar será

$$I = \left[\bar{x} - z_{\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{S}{\sqrt{n}} \right]$$

el cual, para un coeficiente de confianza del 95 % es igual a

$$I = \left[0'753 - 1'96 \sqrt{\frac{0'2686}{35}}, 0'753 + 1'96 \sqrt{\frac{0'2686}{35}} \right] = [0'5813, 0'9247].$$

Si queremos resolver este ejemplo con R, primero introducimos los datos ejecutando (1), un histograma suyo, obtenido ejecutando (2) y que aparece en la Figura 3.1 muestra una fuerte asimetría a la derecha, lo cual sugiere que los valores de actividad no siguen una distribución normal.

```
> x<-c(0.360,1.185,0.524,0.870,0.356,2.567,0.566,
+ 1.789,0.578,0.578,0.892,0.345,0.256,0.987,
+ 0.355,0.989,0.412,0.453,1.987,0.544,0.798,
+ 0.634,0.355,0.455,0.445,0.755,0.423,0.754,
+ 0.452,0.452,0.450,0.511,1.234,0.543,1.501)
```

(1)


```
> hist(x,prob=T)
```

(2)

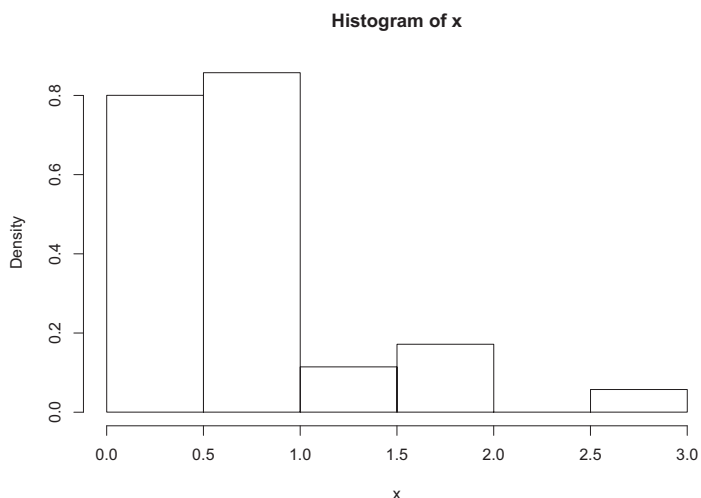


Figura 3.1 : Histograma del Ejemplo 3.3

Si queremos determinar el intervalo de confianza para la media (de una población no necesariamente normal, muestras grandes), de coeficiente de confianza 0'95, ejecutaríamos (3), obteniendo el intervalo en (4).

```
> t.test(x)
```

(3)

```
One Sample t-test
```

```
data: x
t = 8.5953, df = 34, p-value = 4.842e-10
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.5749635 0.9310365
sample estimates:
mean of x
 0.753
```

(4)

El intervalo que obtenemos con R, $[0'5749, 0'9310]$ es algo diferente del que se obtuvo anteriormente debido a que antes se utilizaba la aproximación normal para la determinación de los cuantiles $z_{1-\alpha/2}$ y $z_{\alpha/2}$, mientras que aquí se utilizan los correspondientes de la distribución t de Student. Lo correcto sería lo que hicimos más arriba, pero a medida que n aumenta, apenas habrá diferencia entre ambos.

3.4. Intervalo de confianza para la varianza de una población normal

Dada una muestra aleatoria simple $X_1, \dots, X_n$ de una población $N(\mu, \sigma)$, vamos a determinar el intervalo de confianza para σ^2 , distinguiendo dos casos según sea desconocida o no la media de la población μ .

μ desconocida

El intervalo de confianza buscado será

$$I = \left[\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2} \right]$$

con S^2 la cuasivarianza muestral.

μ conocida

En este caso, el intervalo de confianza será

$$I = \left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n; \alpha/2}^2}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi_{n; 1-\alpha/2}^2} \right].$$

Ejemplo 3.1 (continuación)

Si el terapeuta del Ejemplo 3.1 quiere determinar un intervalo de confianza para la varianza de la variable en estudio, éste será

$$I = \left[\frac{(n-1)S^2}{\chi_{n-1; \alpha/2}^2}, \frac{(n-1)S^2}{\chi_{n-1; 1-\alpha/2}^2} \right]$$

que para un coeficiente de confianza del 99 % proporciona los valores

$$I = \left[\frac{24 \cdot 139}{45'56}, \frac{24 \cdot 139}{9'886} \right] = [73'22, 337'45].$$

Obsérvese que para un tamaño muestral tan pequeño como el que tenemos, el intervalo de confianza al 99 % determinado resulta poco informativo, al tener éste una longitud muy grande.

El correspondiente al 90 %

$$I = \left[\frac{24 \cdot 139}{36'42}, \frac{24 \cdot 139}{13'85} \right] = [91'6, 240'9]$$

tampoco resulta mucho más informativo, perdiendo éste, además, parte del *grado de confianza* que el primero posea. Una de las causas es que, habitualmente, estaremos interesados en estimar la desviación típica y no la varianza, puesto que ésta viene expresada en unidades al cuadrado lo que distorsiona en parte el resultado. El intervalo de confianza para la desviación típica será el de extremos la raíz cuadrada del correspondiente de la varianza. Así por ejemplo, el intervalo correspondiente al 90 % será

$$I = [\sqrt{91'6}, \sqrt{240'9}] = [9'57, 15'52].$$

3.5. Intervalo de confianza para el cociente de varianzas de dos poblaciones normales independientes

Supondremos que $X_1, \dots, X_{n_1}$ e $Y_1, \dots, Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 extraídas respectivamente de dos poblaciones independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

μ_1 y μ_2 conocidas

En este caso, el intervalo de colas iguales es

$$I = \left[\frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2 / \sum_{j=1}^{n_2} (Y_j - \mu_2)^2}{n_1 \cdot F_{n_1, n_2; \alpha/2}}, \frac{n_2 \sum_{i=1}^{n_1} (X_i - \mu_1)^2 / \sum_{j=1}^{n_2} (Y_j - \mu_2)^2}{n_1 \cdot F_{n_1, n_2; 1-\alpha/2}} \right].$$

μ_1 y μ_2 desconocidas

Si las medias poblacionales son desconocidas y las muestras proporcionan cuasivarianzas muestrales S_1^2 y S_2^2 respectivamente, el intervalo de confianza que se obtiene es

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; \alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 1-\alpha/2}} \right].$$

Ejemplo 3.4

Con objeto de estudiar la efectividad de un agente diurético, se eligieron al azar 11 pacientes, aplicando a 6 de ellos dicho fármaco y un placebo a los 5 restantes.

La variable observada en esta experiencia fue la concentración de sodio en la orina a las 24 horas, la cual dio los resultados siguientes:

Diurético :	20'4	62'5	61'3	44'2	11'1	23'7
Placebo :	1'2	6'9	38'7	20'4	17'2	

Supuesto que las concentraciones de sodio, tanto en la población a la que se aplicó el diurético $X_1 \rightsquigarrow N(\mu_1, \sigma_1)$ como a la que se aplicó el placebo $X_2 \rightsquigarrow N(\mu_2, \sigma_2)$, siguen distribuciones normales, en la determinación de un intervalo de confianza para la diferencia de medias poblacionales, veremos que, al ser las muestras pequeñas, necesitamos decidir si las varianzas poblacionales σ_1^2 y σ_2^2 pueden considerarse iguales o no.

Con este propósito se determina un intervalo de confianza para el cociente de dichas varianzas,

$$I = \left[\frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; \alpha/2}}, \frac{S_1^2/S_2^2}{F_{n_1-1, n_2-1; 1-\alpha/2}} \right]$$

que resulta ser, para un coeficiente de confianza del 95 %,

$$I = \left[\frac{483'12/208'52}{9'3645}, \frac{483'12 \cdot 7'3879}{208'52} \right] = [0'247, 17'117]$$

dado que

$$F_{n_1-1, n_2-1; \alpha/2} = F_{5, 4; 0'025} = 9'3645$$

y

$$F_{n_1-1, n_2-1; 1-\alpha/2} = \frac{1}{F_{n_2-1, n_1-1; \alpha/2}} = \frac{1}{F_{4, 5; 0'025}} = \frac{1}{7'3879}.$$

Si queremos resolver este ejemplo con R, primero incorporamos los datos en (1) y (2) y luego ejecutamos (3). El intervalo se obtiene en (4), lógicamente igual al acabado de calcular más arriba.

```
> x<-c(20.4,62.5,61.3,44.2,11.1,23.7) (1)
```

```
> y<-c(1.2,6.9,38.7,20.4,17.2) (2)
```

```
> var.test(x,y) (3)
```

```
F test to compare two variances
```

```
data: x and y
```

```
F = 2.3169, num df = 5, denom df = 4, p-value = 0.4359
```

```
alternative hypothesis: true ratio of variances is not equal to 1
```

```
95 percent confidence interval:
```

```
0.2474174 17.1172392 (4)
```

```
sample estimates:
```

```
ratio of variances
```

```
2.316933
```

Este intervalo de confianza sugiere inferir que el cociente de ambas varianzas poblacionales es 1, es decir, que ambas son iguales, al pertenecer el 1 al intervalo de confianza calculado, razonamiento que justificaremos con detalle en el siguiente capítulo.

El que el 1 parezca estar muy cercano al extremo inferior del intervalo no debe confundirnos ya que la forma de la función de densidad de la F de Snedecor es asimétrica a la derecha por lo que tendrá, en consecuencia, más masa a la izquierda que a la derecha. De hecho, no es un mal ejercicio determinar intervalos de confianza para coeficientes de confianza menores, lo cual acortará la longitud del intervalo de confianza, aunque sensiblemente lo hará más por la derecha que por la izquierda, aunque se observará que éstos siguen conteniendo al 1.

3.6. Intervalo de confianza para la diferencia de medias de dos poblaciones normales independientes

Al igual que en la sección anterior suponemos que $X_1, \dots, X_{n_1}$ e $Y_1, \dots, Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 respectivamente, extraídas de dos poblaciones normales independientes $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$.

σ_1 y σ_2 conocidas

En este caso es

$$\bar{x}_1 - \bar{x}_2 \rightsquigarrow N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

de donde el intervalo de confianza buscado será

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

σ_1 y σ_2 desconocidas. Muestras pequeñas

En esta situación habrá que distinguir según sean

(a) $\sigma_1 = \sigma_2$

En cuyo caso, al ser

$$\frac{\bar{x}_1 - \bar{x}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \rightsquigarrow t_{n_1 + n_2 - 2}$$

obtendremos como intervalo de confianza

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{n_1 + n_2 - 2; \alpha/2} \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

(b) $\sigma_1 \neq \sigma_2$

En este caso, la aproximación de Welch proporciona como intervalo de confianza

$$I = \left[\bar{x}_1 - \bar{x}_2 - t_{f; \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + t_{f; \alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]$$

en donde S_1^2 y S_2^2 son las cuasivarianzas muestrales y f el entero más próximo a

$$\frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1+1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2+1}} - 2$$

Ejemplo 3.4 (continuación)

En la sección anterior concluimos infiriendo que las varianzas poblacionales podían considerarse iguales, admitiendo que las diferencias observadas entre sus estimadores, las cuasivarianzas muestrales, para la muestra concreta que allí se manejaba, era debida al azar y no a que existiera diferencia entre las varianzas poblacionales.

El intervalo de confianza para la diferencia de medias poblacionales $\mu_1 - \mu_2$ será en consecuencia,

$$I = \left[\bar{x}_1 - \bar{x}_2 \mp t_{n_1+n_2-2; \alpha/2} \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right].$$

Utilizando la misma muestra que antes consideramos, práctica muy habitual pero algo más que discutible, obtendríamos el intervalo de confianza, para un coeficiente de confianza del 95 %,

$$I = \left[37'2 - 16'88 \mp 2'262 \sqrt{\frac{5 \cdot 483'12 + 4 \cdot 208'52}{9}} \sqrt{\frac{1}{6} + \frac{1}{5}} \right] = [-5'697, 46'347].$$

Para calcular este intervalo con R, ejecutamos (1) puesto que los datos los habíamos incorporado más arriba. El intervalo se obtiene en (2).

```
> t.test(x,y,var.equal=T) (1)
```

Two Sample t-test

```
data: x and y
t = 1.766, df = 9, p-value = 0.1112
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-5.708955 46.348955 (2)
sample estimates:
mean of x mean of y
37.20      16.88
```

3.7. Intervalo de confianza para la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

Si ahora $X_1, \dots, X_{n_1}$ e $Y_1, \dots, Y_{n_2}$ son dos muestras de tamaños n_1 y n_2 suficientemente grandes, extraídas de dos poblaciones independientes de medias μ_1 y μ_2 respectivamente, de las que sólo suponemos que tienen varianzas σ_1^2 y σ_2^2 finitas, tendremos que

Si σ_1 y σ_2 son conocidas

El intervalo de confianza para $\mu_1 - \mu_2$ con un coeficiente de confianza $1 - \alpha$ es

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right].$$

Si σ_1 y σ_2 son desconocidas

El intervalo de confianza se obtendrá sustituyendo las desconocidas varianzas por las cuasivarianzas muestrales, S_1^2 y S_2^2 , obteniéndose

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$$

Ejemplo 3.5

Los siguientes datos proceden de un estudio del Western Collaborative Group llevado a cabo en California en 1960-1961. En concreto corresponde a 40 individuos de ese estudio de peso elevado, con los que se formaron dos grupos: El Grupo A formado por 20 individuos estresados, ambiciosos y agresivos, y el Grupo B formado por 20 individuos relajados, no competitivos y no estresados. Se midieron en ambos grupos los niveles de colesterol en mgr. por 100 ml. obteniéndose los siguientes datos:

Grupo A:

233 , 291 , 312 , 250 , 246 , 197 , 268 , 224 , 239 , 239
254 , 276 , 234 , 181 , 248 , 252 , 202 , 218 , 212 , 325

Grupo B:

344 , 185 , 263 , 246 , 224 , 212 , 188 , 250 , 148 , 169
226 , 175 , 242 , 252 , 153 , 183 , 137 , 202 , 194 , 213

Vamos a determinar el intervalo de confianza para la diferencia de medias poblacionales con un coeficiente de 0'95. Aunque los tamaños muestrales no son muy grandes, vamos a suponerlos suficientemente grandes para no necesitar la normalidad de las poblaciones de donde proceden los datos.

Como las varianzas poblacionales son desconocidas, el intervalo buscado será

$$I = \left[\bar{x}_1 - \bar{x}_2 - z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}, \bar{x}_1 - \bar{x}_2 + z_{\alpha/2} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right].$$

Con R fácilmente se obtiene el intervalo ejecutando

```
> x1<-c(233,291,312,250,246,197,268,224,239,239,254,276,234,181,248,252,202,218,212,325)
> x2<-c(344,185,263,246,224,212,188,250,148,169,226,175,242,252,153,183,137,202,194,213)
> mean(x1)
[1] 245.05
> mean(x2)
[1] 210.3
> var(x1)
[1] 1342.366
> var(x2)
[1] 2336.747

> mean(x1)-mean(x2)-qnorm(1-0.025)*sqrt(var(x1)/20+var(x2)/20)
[1] 8.166959
> mean(x1)-mean(x2)+qnorm(1-0.025)*sqrt(var(x1)/20+var(x2)/20)
[1] 61.33304
```

Es decir, el intervalo [8'17, 61'33]. Si queremos obtenerlo directamente con R ejecutaríamos

```
> t.test(x1,x2)

Welch Two Sample t-test

data:  x1 and x2
t = 2.5621, df = 35.413, p-value = 0.01481
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 7.227071 62.272929
sample estimates:
mean of x mean of y
 245.05    210.30
```

obteniendo el intervalo [7'22, 62'27].

La pequeña diferencia que se obtiene con el cálculo anterior se debe a que los cálculos con la función `t.test` se hacen con la t de Student, la cual sólo converge a la normal (la que utilizamos en los primeros cálculos) cuando el tamaño muestral es muy grande.

3.8. Intervalos de confianza para datos apareados

En ocasiones nuestros datos $(X_1, Y_1), \dots, (X_n, Y_n)$ tienen una cierta dependencia puesto que miden variables relacionadas, como por ejemplo una variable biomédica observada en los mismos individuos antes X_i y después Y_i de tomar un medicamento. Este tipo de datos recibe el nombre de *datos apareados*.

En estos casos, la forma de actuar consiste en definir la variable unidimensional diferencia $D_i = X_i - Y_i$ y aplicar a sus parámetros los intervalos de confianza antes determinados.

Por ejemplo, si las variables de donde proceden los datos son normales, la variable diferencia D también será normal y si, por ejemplo, las muestras son pequeñas y la varianza es desconocida, el intervalo de confianza para la media $\mu_d = \mu_x - \mu_y$ de coeficiente de confianza $1 - \alpha$, será

$$I = \left[\bar{d} - t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}}, \bar{d} + t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}} \right]$$

en donde es

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n (X_i - Y_i) = \bar{x} - \bar{y} \quad \text{y} \quad S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - Y_i - \bar{d})^2.$$

Ejemplo 3.6

Con objeto de averiguar si la fuerza de la gravedad hace disminuir significativamente la estatura de la personas a lo largo del día, se seleccionaron al azar 10 individuos —mujeres de 25 años—, a las que se midió la estatura (en cm.) por la mañana al levantarse, X_i , y por la noche antes de acostarse, Y_i , obteniéndose los siguientes datos,

X_i	169'7	168'5	165'9	177'8	179'6	168'9	169'2	167'9	181'8	163'3
Y_i	168'2	165'5	164'4	175'7	176'6	166'1	167'1	166'3	179'7	161'5

Si queremos determinar un intervalo de confianza para la diferencia de estaturas medias poblacionales, en primer lugar deberemos calcular las diferencias $D_i = X_i - Y_i$

$$D_i : \quad 1'5 \quad 3 \quad 1'5 \quad 2'1 \quad 3 \quad 2'8 \quad 2'1 \quad 1'6 \quad 2'1 \quad 1'8$$

y como el tamaño muestral es pequeño, $n = 10$, y la varianza poblacional σ_d^2 desconocida, el intervalo de confianza será

$$I = \left[\bar{d} - t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}}, \bar{d} + t_{n-1; \alpha/2} \frac{S_d}{\sqrt{n}} \right]$$

que para un coeficiente de confianza de 0'95 resulta igual a

$$I = \left[2'15 - 2'262 \sqrt{\frac{0'349}{10}}, 2'15 + 2'262 \sqrt{\frac{0'349}{10}} \right] = [1'727, 2'573].$$

Si queremos resolver este ejemplo con R podemos, o bien calcular primero las diferencias $D_i = X_i - Y_i$ y luego ejecutar la función `t.test` a una muestra o, mejor, utilizarla para los pares de datos dados e indicarle que son datos apareados con el argumento `paired`. En concreto, incorporaremos primero los datos en (1) y (2); luego obtenemos un intervalo de confianza de coeficiente de confianza 0'95 ejecutando (3),

```
> x<-c(169.7,168.5,165.9,177.8,179.6,168.9,169.2,167.9,181.8,163.3) (1)
```

```
> y<-c(168.2,165.5,164.4,175.7,176.6,166.1,167.1,166.3,179.7,161.5) (2)
```

```
> t.test(x, y, paired = T) (3)
```

Paired t-test

```
data: x and y
```

```
t = 11.5014, df = 9, p-value = 1.104e-06
```

```
alternative hypothesis: true difference in means is not equal to 0
```

```
95 percent confidence interval:
```

```
1.727125 2.572875 (4)
```

```
sample estimates:
```

```
mean of the differences
```

```
2.15
```

Los resultados aparecen después. Se observa en (4) el intervalo de confianza buscado, idéntico al calculado anteriormente.

Capítulo 4

Contraste de Hipótesis

4.1. Introducción y conceptos fundamentales

Este capítulo es uno de los más importantes del libro ya que los *Contrastes de Hipótesis* son, sin duda alguna, los Métodos Estadísticos más utilizados.

Tanto es así, que el resto de los capítulos del libro son, básicamente, métodos estadísticos basados en contrastes de hipótesis.

Como ilustración de los conceptos que se irán definiendo, supongamos que estamos interesados en averiguar si el consumo habitual de un determinado producto modifica el nivel estándar de colesterol en las personas aparentemente sanas, el cual está fijado en 200 mg/dl. Actualmente parece concluirse que un nivel alto de colesterol es perjudicial en enfermedades cardiovasculares pero que, sin embargo, éste es necesario en la creación de defensas por parte del organismo, por lo que también se consideran perjudiciales niveles bajos de colesterol.

El primer punto a considerar en un contraste de hipótesis es precisamente ése: establecer las *hipótesis* que se quieren contrastar, es decir, comparar.

Así, si en el ejemplo considerado representamos por μ el nivel medio de colesterol en la sangre de las personas que consumen habitualmente el producto en cuestión, el problema que tenemos planteado consiste en decidir si puede admitirse para μ un valor igual a 200 (el producto no modifica el nivel de colesterol) o un valor distinto de 200 (el producto modifica el contenido de colesterol).

Una de las dos hipótesis, generalmente la que corresponde a la situación estándar, recibe el nombre de *hipótesis nula* H_0 , mientras que la otra recibe el nombre de *hipótesis alternativa* H_1 , siendo el *contraste de hipótesis* el proceso de decisión basado en técnicas estadísticas mediante el cual decidimos —inferimos— cuál de las dos hipótesis creemos correcta, aceptándola y rechazando en consecuencia la otra. En este proceso medimos los dos posi-

bles errores que podemos cometer —aceptar H_0 cuando es falsa o rechazar H_0 cuando es cierta— en términos de probabilidades.

Por tanto, nuestro problema se puede plantear diciendo que lo que queremos es realizar el contraste de la hipótesis nula $H_0 : \mu = 200$, frente a la alternativa $H_1 : \mu \neq 200$.

Como todas las técnicas estadísticas, las utilizadas en el contraste de hipótesis se basan en la observación de una muestra, la cual aportará la información necesaria para poder decidir, es decir, para poder contrastar las hipótesis.

Si X representa la variable en observación: nivel de colesterol en la sangre, el contraste de hipótesis concluirá formulando una regla de actuación —denominada también contraste de hipótesis o por no ser excesivamente redundantes, *test de hipótesis* utilizando la terminología anglosajona— la cual estará basada en una muestra de X de tamaño n , $X_1, \dots, X_n$, o más en concreto en una función suya denominada *estadístico del contraste* $T(X_1, \dots, X_n)$, y que habitualmente será una función del estimador natural asociado al parámetro del que se quiere contrastar las hipótesis.

En la realización de un contraste de hipótesis suele ser habitual suponer un modelo probabilístico para la variable X en observación, habitualmente el modelo Normal. Si es posible admitir un modelo se habla de *contrastes paramétricos* que son los que deberemos utilizar siempre que sea posible. A ellos dedicaremos las Secciones 4.2 y 4.4, relajando esta requisito en la Sección 4.3 si el tamaño muestral es grande.

Si no conseguimos ajustar un modelo válido que explique adecuadamente nuestros datos y el tamaño muestral no es grande, deberemos utilizar los denominados *contrastes no paramétricos*, estudiando en la Sección 4.5 el más habitual.

En todo caso, será imprescindible determinar la distribución en el muestreo del estadístico T del test, ya que la filosofía del contraste de hipótesis depende de su distribución en el muestreo, pudiendo formularse de la siguiente forma: si fuera cierta la hipótesis nula H_0 , la muestra, o mejor T , debería de comportarse de una determinada manera —tener una determinada distribución de probabilidad—. Si extraída una muestra al azar, acontece un suceso para T que tenía poca probabilidad de ocurrir si fuera cierta H_0 , —es decir, bajo H_0 —, puede haber ocurrido una de las dos cosas siguientes: o bien es que hemos tenido tan mala suerte de haber elegido una muestra *muy rara* o, lo que es más probable, que la hipótesis nula fuera falsa. La filosofía del contraste de hipótesis consiste en admitir la segunda posibilidad, rechazando en ese caso H_0 , aunque acotando la probabilidad de la primera posibilidad, mediante lo que más adelante denominaremos nivel de significación.

Así en nuestro ejemplo, parece razonable elegir al azar n personas aparentemente sanas a las que, tras haber consumido el producto en cuestión,

mediéramos su nivel de colesterol en sangre, razonando de la siguiente forma: si la hipótesis nula $H_0 : \mu = 200$ fuera cierta, el estimador natural de μ , la media $\bar{x}$ de la muestra obtenida tomaría un valor *cercano* a 200; si, tomada una muestra, este estimador está *lejos* de 200 deberemos rechazar H_0 .

No obstante, los términos *cercano* y *lejano* deben ser entendidos en el sentido de algo con gran probabilidad de ocurrir o poca probabilidad de ocurrir, para lo cual necesitaremos conocer la distribución en el muestreo de T .

Además, estos términos dependen de la magnitud de los errores que estemos dispuestos a admitir, medidos éstos en términos de probabilidades. Puntualicemos estas ideas un poco más.

Errores de tipo I y de tipo II

Para determinar con precisión la regla de actuación en cada caso concreto, debemos considerar los dos errores posibles que podemos cometer al realizar un contraste de hipótesis, los cuales, como antes dijimos, son el de rechazar la hipótesis nula H_0 cuando es cierta, denominado *error de tipo I*, o el de aceptar H_0 cuando es falsa, denominado *error de tipo II*.

Ambos errores son de naturaleza bien distinta; así en el ejemplo considerado, si rechazamos H_0 cuando es cierta, tendremos un coste económico derivado de prohibir un producto no perjudicial, pero si aceptamos H_0 cuando es falsa y permitimos el consumo del producto, pueden producirse graves perjuicios en la salud de los consumidores.

La Estadística Matemática ha deducido tests de hipótesis, es decir reglas de actuación, siguiendo el criterio de fijar una cota superior para la probabilidad de error de tipo I, denominada *nivel de significación*, que maximizan $1 - P\{\text{error de tipo II}\}$, expresión ésta última denominada *potencia del contraste*.

Los tests paramétricos son más potentes que los no paramétricos por lo que son los preferidos, siempre que sea posible admitir un modelo probabilístico válido que los explique

Región crítica y región de aceptación

Los tests de hipótesis, expresados siempre en función de un estadístico T adecuado al problema en cuestión, son de la forma

$$\begin{cases} \text{Aceptar } H_0 & \text{si } T \in C^* \\ \text{Rechazar } H_0 & \text{si } T \in C \end{cases}$$

en donde C y C^* son dos conjuntos disjuntos en los que se ha dividido el conjunto de valores posibles de T . C recibe el nombre de *región crítica* del test, y se corresponde con el conjunto de valores de T en donde se rechaza la hipótesis nula H_0 .

El conjunto complementario, C^* , se denomina *región de aceptación* y se corresponde, como su nombre indica, con el conjunto de valores del estadístico para los cuales se acepta H_0 .

Por completar la terminología propia de los contrastes de hipótesis, diremos que un test es *bilateral* cuando C esté formada por dos intervalos disjuntos y *unilateral* cuando la región crítica sea un intervalo.

Por último, se dice que una hipótesis —nula o alternativa— es *simple* cuando esté formada por un solo valor de parámetro. Si está formada por más de uno, se denomina *compuesta*. Así, el ejemplo considerado se trata de un contraste de hipótesis nula simple —en H_0 está sólo el 200— frente a alternativa compuesta —en H_1 están todos los valores menos el 200.

Siguiendo con el mencionado ejemplo, y denotando $\mu_0 = 200$, hemos dicho que razonablemente deberemos aceptar H_0 cuando $\bar{x}$ esté cerca de μ_0 , Figura 4.1, es decir, cuando sea

$$\mu_0 - c < \bar{x} < \mu_0 + c$$

para un c relativamente pequeño

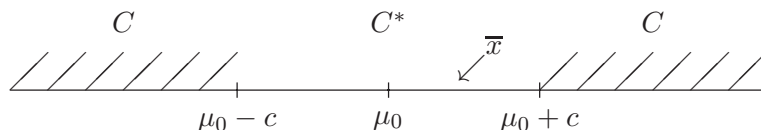


Figura 4.1

o bien, haciendo operaciones, cuando

$$|\bar{x} - \mu_0| < c.$$

Es decir, si $H_0 : \mu = \mu_0$ fuera cierta, cabría esperar que $\bar{x}$ tomara un valor cercano a μ_0 ; en concreto del intervalo $[\mu_0 - c, \mu_0 + c]$, con gran probabilidad, $1 - \alpha$, dependiendo el valor de c de esta probabilidad.

Si observada una muestra concreta, $\bar{x}$ no cae en el intervalo anterior, siguiendo la filosofía del contraste de hipótesis, rechazaremos H_0 , siendo, en consecuencia el mencionado intervalo, la región de aceptación del test.

Determinemos el valor de la constante c : si queremos que la probabilidad de cometer un error de tipo I, es decir, el nivel de significación sea α , deberá ser

$$P\{\bar{x} \in C\} = P\{|\bar{x} - \mu_0| > c\} = \alpha$$

es decir,

$$P\{|\bar{x} - \mu_0| < c\} = 1 - \alpha$$

cuando H_0 es cierta, es decir cuando $\mu = \mu_0$.

Ahora debemos distinguir diversas situaciones. Si podemos admitir un modelo poblacional normal, es decir que $X \rightsquigarrow N(\mu, \sigma)$, sabemos que es

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

con lo que, en la expresión anterior, c deberá ser tal que

$$P\left\{|t_{n-1}| < \frac{c\sqrt{n}}{S}\right\} = 1 - \alpha$$

es decir,

$$c = t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}$$

llevándonos, en definitiva, nuestros razonamientos intuitivos a considerar como test de hipótesis para contrastar a nivel α , $H_0 : \mu = \mu_0$ frente a $H_1 : \mu \neq \mu_0$ el siguiente,

$$\left\{ \begin{array}{ll} \text{Se acepta } H_0 & \text{si } \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1;\alpha/2} \\ \text{Se rechaza } H_0 & \text{si } \frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} > t_{n-1;\alpha/2} \end{array} \right.$$

La Estadística Matemática nos dice que este test es óptimo en el sentido que mencionábamos más arriba.

En concreto, si elegida una muestra aleatoria simple de tamaño $n = 10$ se obtuvo una media muestral $\bar{x} = 202$ y una cuasivarianza muestral de $S^2 = 289$, el contraste $H_0 : \mu = 200$ frente a $H_1 : \mu \neq 200$ lleva a aceptar H_0 a nivel $\alpha = 0'05$ por ser

$$\frac{|202 - 200|}{\sqrt{289/10}} = 0'372 < 2'262 = t_{9;0'025}$$

es decir, a concluir con la no existencia de diferencia significativa a ese nivel.

La deducción exacta de cada contraste óptimo depende de la situación concreta que se tenga: hipótesis de normalidad, muestras grandes, etc., ya que cada una de estas situaciones implica una distribución en el muestreo del estadístico a considerar.

De hecho, la determinación del estadístico a considerar en cada caso —es decir, la forma del contraste— es habitualmente compleja. No obstante, el

lector no debe preocuparse por esta cuestión, de índole matemática, debiendo prestar atención a todo el proceso que un contraste de hipótesis conlleva. Una vez establecido con todo rigor el problema, la elección de la regla óptima será inmediata en los casos considerados en el libro.

Relación entre intervalos de confianza y tests de hipótesis

En el ejemplo anterior, aceptábamos $H_0 : \mu = \mu_0$ cuando

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq t_{n-1;\alpha/2}$$

o bien, haciendo operaciones, cuando

$$\mu_0 \in \left[\bar{x} - t_{n-1;\alpha/2} \frac{S}{\sqrt{n}}, \bar{x} + t_{n-1;\alpha/2} \frac{S}{\sqrt{n}} \right]$$

es decir, cuando la hipótesis nula pertenece al intervalo de confianza correspondiente.

Éste es un hecho bastante frecuente, aunque no una propiedad general, de los contrastes del tipo $H_0 : \theta = \theta_0$ frente a $H_0 : \theta \neq \theta_0$. El intervalo de confianza, de coeficiente de confianza uno menos el nivel de significación, constituye la región de aceptación del test.

Tests de hipótesis unilaterales

Supongamos en el ejemplo antes considerado, que el producto en cuestión es un *snack* elaborado con un determinado aceite. El interés estará entonces centrado en saber si este producto aumenta el nivel medio de colesterol o no. Es decir, en contrastar las hipótesis $H_0 : \mu \leq 200$ frente a $H_1 : \mu > 200$.

Ahora parece claro que la región crítica sea unilateral, Figura 4.2, del tipo $\mu_0 + c$.

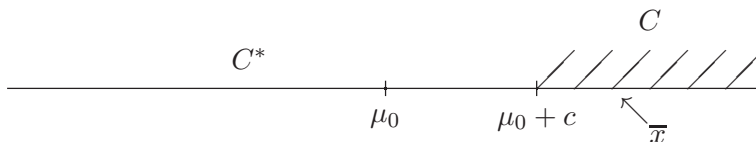


Figura 4.2

Si la probabilidad de error de tipo I es de nuevo α , deberá ser

$$P_{\mu=\mu_0}\{\bar{x} > \mu_0 + c\} = \alpha.$$

Si admitimos la misma situación poblacional anterior, será de nuevo

$$\frac{\bar{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

con lo que en la expresión anterior, c deberá ser tal que

$$P \left\{ t_{n-1} > \frac{c\sqrt{n}}{S} \right\} = \alpha$$

es decir,

$$c = t_{n-1;\alpha} \frac{S}{\sqrt{n}}$$

con lo que se llegaría, en definitiva, a considerar como test de nivel α para contrastar $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$ el siguiente,

$$\begin{cases} \text{Se acepta } H_0 \text{ si } & \frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq t_{n-1;\alpha} \\ \text{Se rechaza } H_0 \text{ si } & \frac{\bar{x} - \mu_0}{S/\sqrt{n}} > t_{n-1;\alpha} \end{cases}$$

En el ejemplo considerado, al ser

$$\frac{202 - 200}{\sqrt{289/10}} = 0'372 < 1'833 = t_{9;0'05}$$

se acepta $H_0 : \mu \leq 200$ al contrastarla frente a $H_1 : \mu > 200$, a nivel $\alpha = 0'05$.

P-valor

Una crítica que puede plantearse el lector respecto a la técnica de los tests de hipótesis, es la dependencia de nuestros resultados en el nivel de significación α elegido antes de efectuar el contraste.

Así surge de forma natural la pregunta: ¿Qué hubiera pasado en el ejemplo anterior si hubiéramos elegido otro α mucho mayor? ¿Se seguiría aceptando H_0 ?

La respuesta evidente es que depende de lo grande que sea α . Si para fijar ideas nos centramos en el contraste unilateral, al ser

$$\frac{\bar{x} - \mu_0}{S/\sqrt{10}} \rightsquigarrow t_9$$

y haber resultado un valor para el estadístico del contraste

$$\frac{\bar{x} - \mu_0}{S/\sqrt{10}} = \frac{202 - 200}{\sqrt{289/10}} = 0'372$$

si hubiéramos elegido por ejemplo $\alpha = 0'4$, hubiéramos rechazado H_0 , ya que $t_{9;0'4} = 0'261 < 0'372$, aunque obsérvese que en este caso la probabilidad de equivocarnos —rechazar H_0 siendo cierta— hubiera sido muy grande, $\alpha = 0'4$.

Parece razonable, por tanto, que independientemente del nivel de significación que hubiéramos elegido, debamos aceptar H_0 , puesto que el nivel de significación más pequeño que hubiéramos tenido que elegir para rechazar H_0 es demasiado grande como para admitir tal probabilidad de error de tipo I.

Este *nivel de significación observado* recibe el nombre de *p-valor* y se define con más precisión como el mínimo nivel de significación necesario para rechazar H_0 .

Obsérvese que al realizar un contraste de hipótesis debemos fijar un nivel de significación antes de tomar la muestra, que habitualmente suele ser $0'1$, $0'05$ ó $0'01$, y para ese nivel de significación elegido, aceptar o rechazar H_0 . Es decir, siempre se llega, por tanto, a una conclusión.

El cálculo del p-valor permite valorar la decisión ya tomada de rechazar o aceptar H_0 , de forma que un p-valor grande —digamos $0'2$ ó más— confirma una decisión de aceptación de H_0 . Tanto más nos lo confirma cuanto mayor sea el p-valor.

Por contra, un p-valor pequeño —digamos $0'01$ ó menos— confirma una decisión de rechazo de H_0 . Tanto más se nos confirmará esta decisión de rechazo cuanto menor sea el p-valor.

En situaciones intermedias, el p-valor no nos indica nada concreto salvo que quizás sería recomendable elegir otra muestra y volver a realizar el contraste.

Si una persona ha tomado una decisión que el p-valor contradice, confirmando éste precisamente la decisión contraria a la adoptada, el individuo lógicamente cambiará su decisión. Por esta razón, muchos de los usuarios de las técnicas estadísticas aplicadas no fijan ya el nivel de significación; simplemente hacen aparecer al final de sus trabajos el p-valor (el cual en muchos paquetes estadísticos se denomina *tail probability*), sacando conclusiones si éste se lo permite o simplemente indicándolo de forma que el lector las saque.

Esta postura, criticable en principio, no lo es más que la de otros investigadores que consideran —por definición— significativo un contraste para un p-valor menor que $0'05$, o la de aquellos otros que sólo contrastan hipótesis a una estrella, dos estrellas o tres estrellas, entendiendo estos niveles de significación, respectivamente como $0'1$, $0'05$ y $0'01$.

En nuestro ejemplo, el p-valor del contraste unilateral será

$$\text{p-valor} = P\{t_9 > 0'372\} = 0'35925$$

y en el bilateral

$$\text{p-valor} = P\{|t_9| > 0'372\} = 2 \cdot P\{t_9 > 0'372\} = 0'7185$$

sugiriendo ambos la aceptación de la hipótesis nula.

Contrastes de Hipótesis con R

Como hemos visto, el intervalo de confianza de un parámetro se corresponde con la región de aceptación de un test de hipótesis bilateral. Por esta razón se utiliza una misma función de R para obtener intervalos de confianza y test de hipótesis sobre un parámetro. En concreto, la función de R que nos va a proporcionar los tests (y los intervalos) es la función `t.test` estudiada brevemente en el capítulo anterior y cuyos argumentos son

```
t.test(x, y = NULL, alternative = "two.sided", mu = 0, paired = FALSE,
      var.equal = FALSE, conf.level = 0.95)
```

Los argumentos `x` e `y` se utilizan para indicar el o los vectores de datos a utilizar en el contraste. El tercer argumento `alternative` presenta tres opciones: `two.sided`, que es la que se utiliza por defecto y que corresponde al caso de contrastes bilaterales; `greater`, correspondiente al caso de hipótesis nula *menor o igual* frente a hipótesis alternativa de *mayor*, y `less` para el caso de hipótesis nula de *mayor o igual* frente a alternativa de *menor*. Deberemos especificar estas opciones entre comillas. Con el argumento `mu` indicamos el valor de la hipótesis nula.

De nuevo `paired` sirve para indicar una situación de datos apareados y `var.equal` si las varianzas poblacionales pueden considerarse o no iguales. El último argumento permite especificar el nivel de significación del test tomándose por defecto el valor 0'05.

4.2. Contraste de hipótesis relativas a la media de una población normal

Supongamos que tenemos una muestra aleatoria simple $X_1, \dots, X_n$ procedente de una población $N(\mu, \sigma)$ y que queremos contrastar hipótesis relativas a la media de la población, μ .

En primer lugar consideraremos el caso de *igual* frente a *distinta*, es decir, el caso en que queremos contrastar si puede admitirse para la media poblacional un determinado valor μ_0 o no.

$\begin{aligned} H_0 : \mu &= \mu_0 \\ H_1 : \mu &\neq \mu_0 \end{aligned}$

En este caso, al igual que ocurre con casi todos los de *igual* frente a *distinta*, la región de aceptación se corresponde con el intervalo de confianza

determinado en el capítulo anterior, aceptándose H_0 cuando y sólo cuando ésta pertenezca al intervalo de confianza.

Así, si suponemos σ **conocida**, fijado un nivel de significación α , aceptaremos $H_0 : \mu = \mu_0$ cuando y sólo cuando

$$\mu_0 \in \left[\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} , \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right]$$

o equivalentemente, haciendo operaciones, cuando

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$$

con lo que podemos concluir diciendo que el test óptimo en esta situación es

• Se acepta H_0 si $\frac{ \bar{x} - \mu_0 }{\sigma/\sqrt{n}} \leq z_{\alpha/2}$ • Se rechaza H_0 si $\frac{ \bar{x} - \mu_0 }{\sigma/\sqrt{n}} > z_{\alpha/2}$

Ejemplo 4.1

Hace 10 años se realizó, en una determinada población, un estudio sobre su estatura cuyo histograma sugirió para dicha variable una distribución normal de media 1'68 m. y desviación típica 6'4 cm.

Ahora se quiere analizar si la estatura media de dicha población ha variado con el tiempo, para lo que se tomó una muestra de tamaño $n = 15$, la cual dio como resultado una media muestral de $\bar{x} = 1'73$ m.

Admitiendo que la distribución modelo sigue siendo normal y que la dispersión en la estatura de dicha población no ha variado en estos diez años, el averiguar si la estatura media de la población se mantiene en los niveles de hace una década o si ha variado significativamente, equivale a contrastar la hipótesis nula $H_0 : \mu = 1'68$ frente a la alternativa $H_1 : \mu \neq 1'68$, en donde μ representa la estatura media poblacional en la actualidad.

Si fijamos un nivel de significación $\alpha = 0'05$, al ser

$$\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} = \frac{|1'73 - 1'68|}{0'064/\sqrt{15}} = 3'026 > 1'96 = z_{0'05/2}$$

debemos rechazar la hipótesis nula H_0 de que la estatura media de la población no ha variado de forma significativa en estos 10 años.

El p-valor del test es

$$P\{|Z| > 3'026\} = 2 \cdot P\{Z > 3'026\} \simeq 0'0025$$

ya que

```
> 2*(1-pnorm(3.026))
[1] 0.002478123
```

Un p-valor tan bajo confirma la decisión tomada.

Si se supone σ **desconocida** el test óptimo en este caso es

• Se acepta H_0 si $\frac{ \bar{x} - \mu_0 }{S/\sqrt{n}} \leq t_{n-1; \alpha/2}$ • Se rechaza H_0 si $\frac{ \bar{x} - \mu_0 }{S/\sqrt{n}} > t_{n-1; \alpha/2}$

a nivel de significación α .

Ejemplo 4.1 (continuación)

Si no se tiene certeza de que la varianza haya permanecido inalterable en los diez años, y la muestra obtenida hubiera dado una cuasivarianza muestral de $0'64 \text{ m}^2$ (la varianza se expresa en unidades al cuadrado), podíamos haber contrastado las hipótesis anteriores, $H_0 : \mu = 1'68$ frente a $H_1 : \mu \neq 1'68$, utilizando un test de la t de Student, que al mismo nivel hubiera aceptado también H_0 al ser

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} = \frac{|1'73 - 1'68|}{\sqrt{0'64/15}} = 0'242 < 2'145 = t_{14; 0'05/2}.$$

El p-valor es

$$\text{p-valor} = 2 \cdot P\{t_{14} > 0'242\} > 2 \cdot P\{t_{14} > 0'258\} = 2 \cdot 0'4 = 0'8$$

ya que

```
> 2*(1-pt(0.258,14))
[1] 0.8001608
```

valor lo suficientemente grande para confirmar la aceptación de H_0 .

$\begin{aligned} H_0 : \mu &\leq \mu_0 \\ H_1 : \mu &> \mu_0 \end{aligned}$

El estudio de los contrastes unilaterales es de suma importancia en el análisis de la efectividad de nuevos productos, donde el aumento de su efectividad ($H_1 : \mu > \mu_0$) o la disminución de alguna característica negativa asociada,

como por ejemplo el tiempo que tarda en hacer efecto ($H_1 : \mu < \mu_0$) son las hipótesis de interés.

En estos casos, el objetivo es rechazar H_0 con un p-valor pequeño, lo que conduce a quedarnos con la hipótesis de interés H_1 , con un error pequeño en la inferencia, el error de rechazar H_0 siendo cierta, error suministrado por el p-valor.

La distribución en el muestreo de $\bar{x}$ en los supuestos que se establecen, así como las consideraciones hechas al hablar de las hipótesis unilaterales, llevan a la Estadística Matemática a proponer como test óptimo para contrastar $H_0 : \mu \leq \mu_0$ frente a $H_1 : \mu > \mu_0$,

Si σ es conocida

El test óptimo indica que

• Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$ • Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

Si σ es desconocida

En este caso, el test óptimo indica que

• Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq t_{n-1;\alpha}$ • Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} > t_{n-1;\alpha}$

Ejemplo 4.2

Un laboratorio farmacéutico piensa que un nuevo medicamento fabricado por ellos prolonga significativamente la vida de los enfermos de SIDA, establecida en la actualidad en una media de dos años desde que la enfermedad se manifiesta.

Con objeto de validar su nuevo producto, y admitiendo que el tiempo de vida sigue una distribución normal de media μ , el laboratorio contrastó la hipótesis nula $H_0 : \mu \leq 2$ frente a la alternativa $H_1 : \mu > 2$, utilizando una muestra aleatoria de $n = 18$ pacientes, la cual le proporcionó una media de $\bar{x} = 2'8$ años y una cuasidesviación típica muestral de $S = 1'2$ años. Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{2'8 - 2}{1'2/\sqrt{18}} = 2'8284$$

el laboratorio rechazaría H_0 —validando en consecuencia su producto— con un p-valor suficientemente pequeño, aproximadamente igual a 0'006 ya que

```
> 1-pt(2.8284,17)
[1] 0.005795382
```

$$\begin{aligned} H_0 : \mu &\geq \mu_0 \\ H_1 : \mu &< \mu_0 \end{aligned}$$

Los mismos razonamientos anteriores llevan a proponer los siguientes tests para las hipótesis simétricas aquí consideradas.

Si σ es conocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$

Si σ es desconocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \geq t_{n-1;1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} < t_{n-1;1-\alpha}$

Ejemplo 4.3

La rapidez con la que un determinado medicamento actúa es esencial ante infartos agudos de miocardio. Los fármacos que se administran en la actualidad tardan en actuar una media de 30 segundos.

Un laboratorio afirma que el producto recién elaborado por ellos, actúa en menos tiempo. ¿Podemos recomendar su utilización?

El contraste de hipótesis que se plantea es $H_0 : \mu \geq 30$ frente a $H_1 : \mu < 30$. Si una muestra de $n = 10$ pacientes dio un tiempo medio de reacción de 28 segundos y una cuasivarianza de $S^2 = 16$ segundos al cuadrado, no podemos rechazar H_0 a nivel $\alpha = 0'05$ ya que

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{28 - 30}{4/\sqrt{10}} = -1'58 > -1'833 = t_{9,0'95}$$

al ser

```
> qt(0.05,9)
[1] -1.833113
```

El p-valor del test es

```
> pt(-1.58,9)
[1] 0.07428219
```

no es concluyente aunque podemos concluir afirmando que no existen evidencias claras de la efectividad del nuevo producto *al nivel de significación indicado*.

4.3. Contraste de hipótesis relativas a la media de una población no necesariamente normal. Muestras grandes

La obtención de tamaños muestrales suficientemente grandes —digamos mayores de 30— evita la obligación de suponer normalidad en la distribución modelo, alcanzándose, no obstante, resultados análogos a cuando se verifica tal suposición.

La normalidad en la distribución asintótica de $\bar{x}$, añade la peculiaridad de hacer que los puntos críticos sean ahora abscisas de normales estándar, tanto si la varianza poblacional es conocida como si no lo es.

Población no necesariamente normal

Supongamos que $X_1, \dots, X_n$ es una muestra aleatoria simple de tamaño suficientemente grande como para poder admitir como distribución asintótica de $\bar{x}$ la siguiente,

$$\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

En este caso, considerando los tres tipos de tests y distinguiendo, de nuevo, la situación en la que la varianza es conocida y la situación en la que es desconocida, tenemos los siguientes contrastes,

$$\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}$$

σ conocida

El test óptimo que se propone es la siguiente regla de actuación

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$

σ desconocida

Si σ es desconocida, entonces el test óptimo es

- Se acepta H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} > z_{\alpha/2}$

Ejemplo 4.4

Un grupo de arqueólogos considera que la capacidad craneal es el factor determinante en la clasificación de restos humanos del paleolítico, variable que se admite sigue una distribución normal. En concreto, una capacidad craneal de 1500 cm^3 lleva a clasificar a un esqueleto como de raza *Neanderthal*.

Ante el hallazgo de 8 esqueletos en una necrópolis de la mencionada época, los arqueólogos calcularon una capacidad craneal media en dichos restos de 1450 cm^3 y una desviación típica muestral de 10 cm^3 .

En estas condiciones, la determinación de si los restos hallados pueden considerarse como de raza *Neanderthal* puede conseguirse contrastando la hipótesis nula $H_0 : \mu = 1500$ frente a $H_1 : \mu \neq 1500$ en donde μ representa la capacidad craneal media de la población de restos encontrados. Como es

$$\frac{|\bar{x} - \mu_0|}{S/\sqrt{n}} = \frac{|1450 - 1500|}{10'69/\sqrt{8}} = 13'23$$

y el p-valor del test

```
> 2*(1-pnorm(13.23))
[1] 0
```

prácticamente cero, la conclusión que puede sacarse es que claramente los restos no eran de raza *Neanderthal*.

$$\begin{array}{l} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}$$

Si σ es conocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha$

Si σ es desconocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \leq z_\alpha$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} > z_\alpha$

Ejemplo 4.5

En una muestra de 49 adolescentes que sirvieron de sujetos en un estudio inmunológico, una variable de interés fue el diámetro de reacción en la piel ante un antígeno. La media y la desviación típica muestrales fueron 39 y 11 mm. respectivamente.

Si la reacción media habitual es de 30 mm. cabe preguntarse si la reacción observada fue mayor de lo esperado. Es decir, parece razonable contrastar la hipótesis nula $H_0 : \mu \leq 30$ frente a la alternativa $H_1 : \mu > 30$.

Obsérvese que no tiene sentido plantearse el contraste de las hipótesis complementarias $H_0 : \mu \geq 30$ frente $H_1 : \mu < 30$, ya que éste tiene como región crítica la cola de la izquierda y, al haberse observado una media muestral mayor que la hipótesis nula, siempre se aceptaría H_0 . Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{39 - 30}{11'114/\sqrt{49}} = 5'6685 > 1'645 = z_{0'05}$$

rechazaremos la hipótesis nula a nivel $\alpha = 0'05$. El p-valor

```
> 1-pnorm(5.6685)
[1] 7.202654e-09
```

confirma, fuertemente, esta decisión.

$$\begin{array}{l} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}$$

Si σ es conocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < z_{1-\alpha}$

Si σ es desconocida

- Se acepta H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} \geq z_{1-\alpha}$
- Se rechaza H_0 si $\frac{\bar{x} - \mu_0}{S/\sqrt{n}} < z_{1-\alpha}$

Ejemplo 4.6

Un grupo de historiadores norteamericanos está interesado en averiguar si la edad media de los soldados de la Unión en la época previa a la guerra civil americana de 1861 era menor de 30 años.

Con este propósito el grupo consideró *Fort Moultrie*, en Carolina del Sur, suficientemente representativo de los 75 fuertes con los que contaba Estados Unidos en 1850, eligiendo de allí una muestra de tamaño $n = 45$ para la que se obtuvo, según el Censo de Carolina del Sur de 1850, una media de $\bar{x} = 28'3$ años y una cuasidesviación típica $S = 5'96$.

Planteando el contraste de las hipótesis $H_0 : \mu \geq 30$ frente a $H_1 : \mu < 30$ y dado que el tamaño muestral es suficientemente grande, la suposición de normalidad para la variable edad no es requerida. Como es

$$\frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{28'3 - 30}{5'96/\sqrt{45}} = -1'91 < -1'645 = z_{1-0'05}$$

podemos rechazar H_0 a nivel $\alpha = 0'05$, infiriendo, por tanto, una edad significativamente inferior a 30 años en los soldados, aunque con un p-valor,

$$P\{Z < -1'91\} = 0'0281$$

ya que es

```
> pnorm(-1.91)
[1] 0.02806661
```

el cual no es concluyente.

4.4. Contraste de hipótesis relativas a la varianza de una población normal

En toda la sección supondremos que tenemos una muestra $X_1, \dots, X_n$ de una población normal $N(\mu, \sigma)$ y que estamos interesados en realizar contrastes sobre la varianza de dicha distribución.

Apuntemos, además, que las hipótesis referentes a la desviación típica se contrastarían utilizando las raíces cuadradas de los tests que aparecen a continuación.

$H_0 : \sigma^2 = \sigma_0^2$ $H_1 : \sigma^2 \neq \sigma_0^2$
--

μ conocida

Si la media es conocida, el test óptimo a utilizar de nivel de significación α , es

$$\begin{aligned}
&\bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \in \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \\
&\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \notin \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right]
\end{aligned}$$

μ desconocida

En este caso la regla a utilizar será

$$\begin{aligned}
&\bullet \text{ Se acepta } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} \in \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] \\
&\bullet \text{ Se rechaza } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} \notin \left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right]
\end{aligned}$$

Ejemplo 4.7

Se realizó un experimento con objeto de analizar la destreza de 18 estudiantes de enfermería, observando en ellos una medida de la destreza manual, la cual dio una cuasivarianza muestral de $S^2 = 1349$.

Supuesto que esta medida de la destreza sigue una distribución normal, ¿puede concluirse que la varianza poblacional es diferente de 2600, a nivel $\alpha = 0'05$?

Al no suponerse la media poblacional conocida, utilizaremos el segundo test. Como es

$$\left[\chi_{n-1; 1-\frac{\alpha}{2}}^2, \chi_{n-1; \frac{\alpha}{2}}^2 \right] = \left[\chi_{17; 1-0'025}^2, \chi_{17; 0'025}^2 \right] = [7'564, 30'19]$$

y

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{17 \cdot 1349}{2600} = 8'82 \in [7'564, 30'19]$$

no podemos rechazar H_0 a ese nivel. El p-valor será

```
> 2*(pchisq(8.82,17))
[1] 0.10852
```

bastante claro en la aceptación de la hipótesis nula.

$$\begin{array}{l} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}$$

μ conocida

En este caso el test óptimo es

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \leq \chi_{n;\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} > \chi_{n;\alpha}^2 \end{array}$$

μ desconocida

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} \leq \chi_{n-1;\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{(n-1)S^2}{\sigma_0^2} > \chi_{n-1;\alpha}^2 \end{array}$$

Ejemplo 4.8

Con objeto de estudiar la cantidad de proteínas contenidas en el líquido amniótico, se seleccionaron al azar 16 mujeres embarazadas, obteniéndose una cuasidesviación típica muestral de $S = 0'7$ gramos por cada 100 ml. Admitiendo normalidad en dicha variable, contrastar, a nivel $0'05$, si la desviación típica poblacional puede considerarse mayor que $0'6$.

Como es $\chi_{15;0'05}^2 = 25$ y

$$\frac{S \sqrt{n-1}}{\sigma_0} = \frac{0'7 \sqrt{15}}{0'6} = 4'518 < 5$$

se acepta $H_0 : \sigma \leq 0'6$. El p-valor será

$$P \left\{ \sqrt{\chi_{15}^2} > 4'518 \right\} = P \left\{ \chi_{15}^2 > 20'41 \right\} = 0'157$$

ya que

```
> 1-pchisq(20.41,15)
[1] 0.1567623
```

bastante claro en la aceptación de H_0 .

$$\begin{aligned} H_0 : \sigma^2 &\geq \sigma_0^2 \\ H_1 : \sigma^2 &< \sigma_0^2 \end{aligned}$$

μ conocida

En esta situación, el test óptimo indica que

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} \geq \chi_{n-1;1-\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{\sum_{i=1}^n (X_i - \mu)^2}{\sigma_0^2} < \chi_{n-1;1-\alpha}^2 \end{aligned}$$

μ desconocida

$$\begin{aligned} \bullet \text{ Se acepta } H_0 \text{ si } & \frac{(n-1)S^2}{\sigma_0^2} \geq \chi_{n-1;1-\alpha}^2 \\ \bullet \text{ Se rechaza } H_0 \text{ si } & \frac{(n-1)S^2}{\sigma_0^2} < \chi_{n-1;1-\alpha}^2 \end{aligned}$$

Ejemplo 4.9

Los pesos de 30 bebés recién nacidos que habían sido elegidos al azar, dieron una cuasidesviación típica muestral de 165 gramos. Admitiendo que los pesos en los recién nacidos siguen una distribución normal, contrastar las hipótesis $H_0 : \sigma^2 \geq 32000$ frente a $H_1 : \sigma^2 < 32000$, a nivel $\alpha = 0'05$.

Como es $\chi_{n-1;1-\alpha}^2 = \chi_{29;0'95}^2 = 17'71$ y

$$\frac{(n-1)S^2}{\sigma_0^2} = \frac{29 \cdot 165^2}{32000} = 24'67 > 17'71$$

se acepta H_0 . Además, el p-valor = $P\{\chi_{29}^2 < 24'67\} \simeq 0'3$ ya que

```
> pchisq(24.67, 29)
[1] 0.3047471
```

confirma esta decisión.

4.5. El contraste de los rangos signados de Wilcoxon

Si no podemos admitir un modelo normal para los datos observados y el tamaño de la muestra no es grande, debemos utilizar un test no paramétrico. En el caso de considerar sólo una población, el test más utilizado es el *contraste de los rangos signados de Wilcoxon*.

La idea es la misma de los tests paramétricos acabados de estudiar, analizando si puede admitirse un valor para la media de la distribución de la variable en estudio puesto que, como ya comentamos anteriormente, ésta viene representada por su media.

En los contrastes no paramétricos, como el que aquí estudiaremos, la distribución de la variable en estudio se representa por su mediana M , siendo éste el parámetro al que nos referiremos en las hipótesis a contrastar.

$H_0 : M = M_0$ $H_1 : M \neq M_0$

Aunque este test lo ejecutaremos con R, por comentar la razón de su definición, si $X_1, \dots, X_n$ es una muestra aleatoria de la variable en observación y $D_i = X_i - M_0$ las diferencias de la muestra con la mediana a contrastar M_0 , primero se ordenarían sus valores absolutos $|D_1|, \dots, |D_n|$ asignando a cada uno su rango $r(|D_i|)$, es decir, al menor $|D_i|$ el valor 1 y así hasta el último al que asignamos el valor n , utilizando en el test de Wilcoxon como estadístico de contraste, T^+ , la *suma de los rangos de las diferencias positivas*.

Contraste de hipótesis

Valores muy grandes o muy pequeños de T^+ desacreditarán la hipótesis nula $H_0 : M = M_0$ en favor de la alternativa $H_1 : M \neq M_0$, con lo que fijado un nivel de significación α ,

- Se acepta H_0 si $\frac{n(n+1)}{2} - t_{\alpha/2} < T^+ < t_{\alpha/2}$
- Se rechaza H_0 si $T^+ \leq \frac{n(n+1)}{2} - t_{\alpha/2}$ ó $T^+ \geq t_{\alpha/2}$

en donde $t_{\alpha/2}$ es el punto crítico tal que $P\{T^+ \geq t_{\alpha/2}\} = \alpha/2$.

Contraste de los rangos signados de Wilcoxon con R

El test de los rangos signados de Wilcoxon se ejecuta con la función

```
wilcox.test(x,alternative="two.sided",mu=0)
```

en donde incluiremos en el primer argumento `x` el vector de observaciones. Con el argumento `alternative` podemos elegir el tipo de test que vamos a ejecutar, bilateral (que es el que se utiliza por defecto), `less` o `greater` si la hipótesis alternativa que queremos contrastar es, respectivamente, menor o mayor. Con `mu` podemos señalar el valor de la hipótesis a contrastar, eligiendo la función el valor 0 por defecto.

Si hay observaciones iguales a la hipótesis a contrastar deberemos eliminarlas, reduciendo el tamaño muestral, o promediarlas. El ordenador nos avisará si aparecen empates entre los valores absolutos de las diferencias a ordenar por rangos aunque no las elimina sino que las promedia.

Ejemplo 4.10

Se está llevando a cabo un experimento con objeto de medir los efectos que produce la inhalación prolongada de óxido de cadmio.

Los niveles de hemoglobina, en gramos, de cuatro ratones elegidos al azar de un laboratorio en donde existe la contaminación en estudio fueron 14'4, 15'9, 13'8, 15'3. ¿Puede admitirse la hipótesis nula de un promedio poblacional de 15 gramos?

Como con 4 datos suponer un modelo normal es muy aventurado, utilizaremos el test de los rangos signados de Wilcoxon para contrastar $H_0 : M = 15$ frente a $H_1 : M \neq 15$.

Para ello, después de incorporar los datos en (1), ejecutamos (2) para obtener en (3) el valor del estadístico $T^+ = 4$ y el p-valor, 0'875, suficientemente grande como para aceptar la hipótesis nula.

```
> x<-c(14.4,15.9,13.8,15.3) (1)
```

```
> wilcox.test(x,mu=15) (2)
```

```
Wilcoxon signed rank test
```

```
data: x
```

```
V = 4, p-value = 0.875 (3)
```

```
alternative hypothesis: true location is not equal to 15
```

$$\begin{array}{l} H_0 : M \leq M_0 \\ H_1 : M > M_0 \end{array}$$

En este caso, fijado un nivel de significación α

- Se acepta H_0 si $T^+ < t_\alpha$
- Se rechaza H_0 si $T^+ \geq t_\alpha$

en donde de nuevo t_α es el menor número entero tal que

$$P\{T^+ \geq t_\alpha\} \leq \alpha.$$

Ejemplo 4.11

Se realizó un estudio con objeto de averiguar si el número de linfocitos en los animales de laboratorio era mayor de 2500 por milímetro cúbico.

Para ello se seleccionaron al azar 15 de dichos animales para los que se obtuvieron los siguientes datos sobre su número de linfocitos, expresados en miles por milímetro cúbico

Animal	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Linfo.	2'3	2'9	1'6	2	4'2	3'1	2'3	2'5	2	1'6	3'3	4'1	4	3	2'8

Las hipótesis a contrastar son $H_0 : M \leq 2'5$ frente a $H_1 : M > 2'5$.

Al haberse observado un valor igual a 2'5, lo ignoraremos. Para resolver este ejemplo con R, primero incorporamos los datos en (1), puesto que no los habíamos incluido antes al ejecutar el test de los signos. Recordemos que este test no tiene en cuenta el valor de las observaciones; sólo si son mayores o menores que la hipótesis a contrastar. En (2) ejecutamos el test de Wilcoxon, calculando el valor aproximado del p-valor y sin corrección de continuidad.

```
> x<-c(2.3,2.9,1.6,2,4.2,3.1,2.3,2,1.6,3.3,4.1,4,3,2.8) (1)
```

```
> wilcox.test(x,alternative="greater",mu=2.5) (2)
```

Wilcoxon signed rank test with continuity correction

```
data: x
V = 69, p-value = 0.1572 (3)
alternative hypothesis: true location is greater than 2.5
```

En (3) obtenemos el valor del estadístico del test, $V=69$, y el valor aproximado del p-valor, 0'1498, suficientemente grande como para aceptar la hipótesis nula y concluir que no puede admitirse un promedio para el número de linfocitos en los animales de laboratorio sea mayor de 2500 por milímetro cúbico.

$$\begin{array}{l} H_0 : M \geq M_0 \\ H_1 : M < M_0 \end{array}$$

Para este último contraste unilateral, fijado un nivel de significación α

- Se acepta H_0 si $T^+ > \frac{n(n+1)}{2} - t_\alpha$
- Se rechaza H_0 si $T^+ \leq \frac{n(n+1)}{2} - t_\alpha$

siendo de nuevo t_α el menor número entero tal que

$$P\{T^+ \geq t_\alpha\} \leq \alpha.$$

Capítulo 5

Comparación de Poblaciones

5.1. Introducción

En Estadística Aplicada es habitual la *Comparación de Poblaciones* es decir, la comparación de dos o más grupos de datos con objeto de analizar, mediante un contraste de hipótesis, si estos conjuntos de datos pueden considerarse iguales o si, por ejemplo en la comparación de dos grupos de datos, uno de ellos procedente de las observaciones de un nuevo medicamento, puede considerarse mejor que el otro.

Los tests utilizados en la Comparación de Poblaciones se pueden clasificar en dos grandes grupos: *Tests Paramétricos*, que requieren de la normalidad de los datos, es decir, que pueda admitirse que las observaciones proceden de un modelo normal y *Tests no Paramétricos* que no exigen esta suposición.

Dentro de los *Tests Paramétricos* hay que distinguir si puede admitirse que las varianzas de las poblaciones a comparar son iguales (suposición de *homocedasticidad*) y si no puede admitirse este requisito.

Si las muestras son suficientemente grandes, estos requisitos se relajan y pueden utilizarse estos tests.

Si los tamaños muestrales son pequeños y no se verifican las suposiciones necesarias para poder ser utilizados, es necesario ejecutar *Tests no Paramétricos* como el de Wilcoxon-Mann-Whitney en la comparación de dos poblaciones o el de Kruskal-Wallis en la comparación de más de dos poblaciones. Esto en el caso de que tengamos observaciones de alguna variable de tipo cuantitativo ya que si sólo tenemos recuentos de observaciones, deberemos utilizar el test de la χ^2 de Homogeneidad de Varias Muestras.

Estas diferencias se resumen en el cuadro que sigue para la comparación de dos poblaciones:

•Tests Paramétricos	$\left\{ \begin{array}{l} \text{Muestras pequeñas} \left\{ \begin{array}{l} \text{Varianzas iguales: Test de la } t \text{ de Student (5.5)} \\ \text{Varianzas distintas: Test de Welch (5.5)} \end{array} \right. \\ \text{Muestras grandes: Tests basados en la normal (5.6)} \end{array} \right.$
•Tests no Paramétricos	$\left\{ \begin{array}{l} \text{Observaciones de una variable: Wilcoxon-Mann-Whitney (5.7)} \\ \text{Recuentos de observaciones: Test } \chi^2 \text{ de homogeneidad (5.10)} \end{array} \right.$

mientras que en el caso de la comparación de más de dos poblaciones, la situación sería la siguiente:

•Tests Paramétricos	$\left\{ \begin{array}{l} \text{Muestras pequeñas} \left\{ \begin{array}{l} \text{Varianzas iguales: ANOVA (5.8)} \\ \text{Varianzas distintas: Test de Welch (5.8)} \end{array} \right. \\ \text{Muestras grandes: Test de Welch (5.8)} \end{array} \right.$
•Tests no Paramétricos	$\left\{ \begin{array}{l} \text{Rangos de observaciones: Kruskal-Wallis (5.9)} \\ \text{Recuentos de observaciones: Test } \chi^2 \text{ de homogeneidad (5.10)} \end{array} \right.$

Entre paréntesis aparece la sección en la que se estudia cada test, alguno de los cuales es el mismo tanto para comparar dos poblaciones como más de dos.

Son mejores, es decir, más potentes, los tests paramétricos por lo que siempre que podamos serán estos tests los que debamos ejecutar. Un poco más abajo estudiaremos la posibilidad de transformar los datos para que se cumplan las suposiciones necesarias y poder utilizar tests paramétricos para los datos transformados. Hay una última posibilidad que se sale de los objetivos de este libro; se trata de utilizar Métodos Estadísticos Robustos. Aquellos lectores interesados en este tipo de técnicas puede leer el libro del autor de este texto, *Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo*.

En los tests paramétricos, las poblaciones a comparar vienen representadas por sus medias por lo que dichos tests harán referencia a ellas mientras que en los tests no paramétricos, serán las medianas los parámetros a contrastar, excepto en el de la χ^2 en donde la hipótesis nula será, sencillamente, la homogeneidad de las poblaciones.

Los tests de comparación de más de dos poblaciones reciben habitualmente el nombre de tests de Análisis de la Varianza ANOVA.

Dado que las suposiciones que deben verificar los datos es un requisito previo en la elección del test a utilizar, comenzaremos el capítulo con los análisis de normalidad y homocedasticidad de los datos. Ambas suposiciones pueden ser comprobadas gráficamente y, mejor aún, mediante un test de hipótesis.

5.2. Análisis de la Normalidad

El Análisis de la Normalidad de unos datos se puede efectuar gráficamente con ayuda del denominado *Gráfico de normalidad* o *qq-plot* el cual consiste en representar en el eje de abscisas los cuantiles de la normal estándar y en el eje de ordenadas los cuantiles de la muestra; si estos pares de puntos están más o menos en la diagonal del gráfico, se tendrá que los cuantiles muestrales serán similares a los de la $N(0, 1)$ y podremos concluir con la normalidad de los datos. Este gráfico se puede obtener fácilmente con R gracias a la función `qqnorm`.

Obtendremos también el diagrama de hojas y ramas, que vimos en el Capítulo 1 que se podría conseguir con la función `stem` para completar el Análisis de Normalidad.

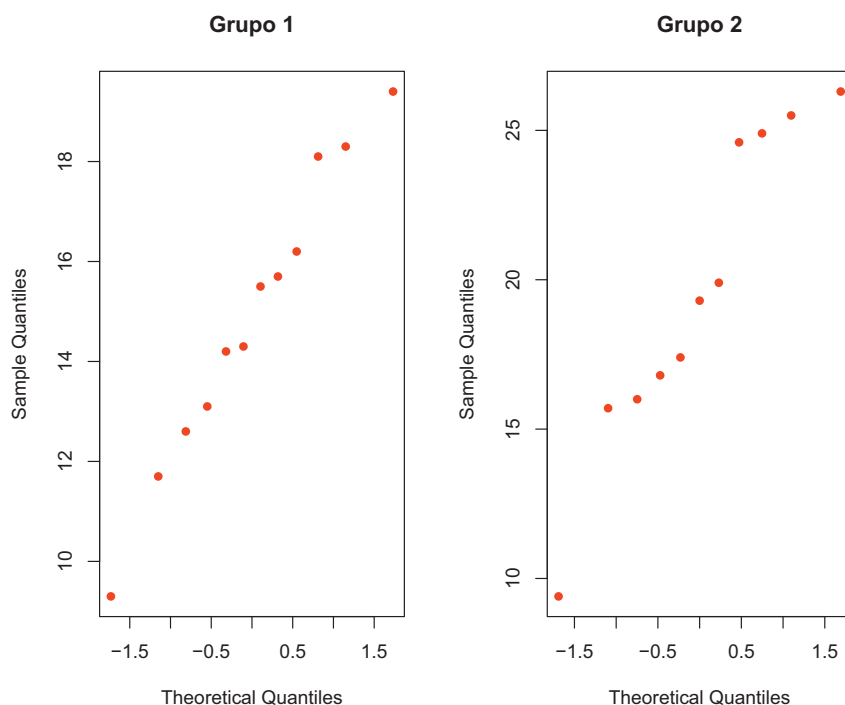


Figura 5.1 : qq-plots del Ejemplo 5.1

Ejemplo 5.1

Un grupo de científicos de una estación antártica, estuvo de acuerdo en participar en un estudio nutricional el cual se proponía analizar los niveles de vitamina C en personas que viven en un clima extremadamente frío.

Con este objetivo, las personas de la estación fueron divididas al azar en dos grupos. Al Grupo 1 le fue administrado un suplemento de vitamina C y el Grupo 2 fue utilizado como grupo control.

Los datos de los niveles, en $\mu\text{g}/10^8$ células, de ácido ascórbico en sangre fueron (Fuente: Dr. P. Gormley, Antarctic Division, Australian Department of Science and Technology)

Grupo 1	18'3	9'3	12'6	15'7	14'2	13'1	14'3	16'2	18'1	19'4	15'5	11'7
Grupo 2	24'9	16	26'3	25'5	19'3	16'8	15'7	24'6	19'9	9'4	17'4	

Después de incorporar los datos podemos conseguir el qq-plot ejecutando la siguiente secuencia de instrucciones con la que obtenemos la Figura 5.1. La normalidad suministrada por el qq-plot del Grupo 1 parece clara pero la del Grupo 2 no parece tan clara.

```
> Grupo1<-c(18.3,9.3,12.6,15.7,14.2,13.1,14.3,16.2,18.1,19.4,15.5,11.7)
> Grupo2<-c(24.9,16,26.3,25.5,19.3,16.8,15.7,24.6,19.9,9.4,17.4)

> par(mfrow=c(1,2))
> qqnorm(Grupo1,pch=16,col=2,main="Grupo 1")
> qqnorm(Grupo2,pch=16,col=2,main="Grupo 2")
```

Si obtenemos el gráfico de hojas y ramas de ambos grupos,

```
> stem(Grupo1,scale=2)

The decimal point is at the |

 8 | 3
10 | 7
12 | 61
14 | 2357
16 | 2
18 | 134

> stem(Grupo2)

The decimal point is 1 digit(s) to the right of the |

0 | 9
1 |
1 | 66779
2 | 0
2 | 5566
```

las conclusiones tampoco son claras, especialmente si movemos la escala con el argumento **scale**. Ésta es la razón principal por la que no es bueno sacar conclusiones con gráficos: un cambio en la escala permite obtener conclusiones diferentes. Siempre será preferible un test de hipótesis que permite valorar la probabilidad de error mediante el p-valor.

Básicamente hay dos tests de hipótesis para contrastar la normalidad: el test de Kolmogorov-Smirnov que es potente para tamaños muestrales grandes, pero cuando éstos son pequeños, el test de Kolmogorov-Smirnov tiende a ser conservador, es decir, a aceptar la hipótesis nula, por lo que se recomienda utilizar el test de Shapiro-Wilk, seguramente el test más potente en detectar la no normalidad de unos datos. El primer test para ambas poblaciones se obtiene ejecutando

```
> ks.test(Grupo1,"pnorm",mean(Grupo1),sd(Grupo1))
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: Grupo1
D = 0.1135, p-value = 0.9929
alternative hypothesis: two-sided
```

```
> ks.test(Grupo2,"pnorm",mean(Grupo2),sd(Grupo2))
```

```
One-sample Kolmogorov-Smirnov test
```

```
data: Grupo2
D = 0.1913, p-value = 0.7489
alternative hypothesis: two-sided
```

que claramente acepta la normalidad con p-valores 0'9929 y 0'7489. Los tests de Shapiro-Wilk, los ejecutaremos con

```
> shapiro.test(Grupo1)
```

```
Shapiro-Wilk normality test
```

```
data: Grupo1
W = 0.9794, p-value = 0.9811
```

```
> shapiro.test(Grupo2)
```

```
Shapiro-Wilk normality test
```

```
data: Grupo2
W = 0.9233, p-value = 0.3468
```

que también terminan aceptándola pero, como vemos, con menos contundencia.

5.3. Análisis de la Homocedasticidad

El Análisis de la homocedasticidad se puede hacer gráficamente mediante un *Gráfico de cajas*, obtenido con la función `boxplot`.

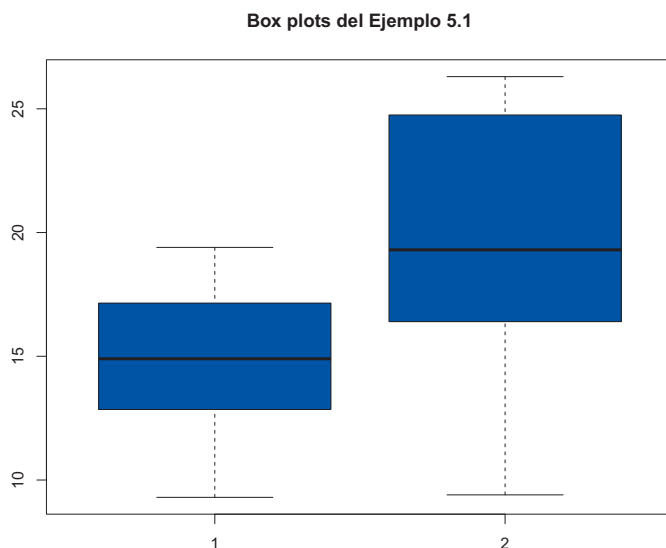


Figura 5.2 : Gráficos de cajas del Ejemplo 5.1

Ejemplo 5.1 (continuación)

Si queremos obtener el gráfico de cajas de los datos ejecutaríamos los comandos

```
> grupo<-c(rep("1",12),rep("2",11))  
> acido<-c(Grupo1,Grupo2)  
> boxplot(acido~grupo,col=4)
```

obteniendo la Figura 5.2 que parece mostrar mayor varianza en el Grupo 2.

Para concluir o no con la igualdad de las varianzas de ambos grupos es mejor ejecutar un test de hipótesis. En el caso de sólo dos poblaciones podemos contrastar las hipótesis $H_0 : \sigma_1^2 = \sigma_2^2$ frente a $H_1 : \sigma_1^2 \neq \sigma_2^2$ en el caso de que se admita normalidad de los datos (lógicamente con medias desconocidas) mediante el correspondiente intervalo de confianza

- Se acepta H_0 si $\frac{S_1^2}{S_2^2} \in \left[F_{n_1-1, n_2-1; 1-\frac{\alpha}{2}}, F_{n_1-1, n_2-1; \frac{\alpha}{2}} \right]$
- Se rechaza H_0 si $\frac{S_1^2}{S_2^2} \notin \left[F_{n_1-1, n_2-1; 1-\frac{\alpha}{2}}, F_{n_1-1, n_2-1; \frac{\alpha}{2}} \right]$

que con R se ejecuta

```
var.test(x, y, ratio, alternative="two.sided", conf.level = 0.95)
```

en donde incorporamos los datos en los argumentos `x` e `y`. En `ratio` especificamos la hipótesis nula, que será `ratio = 1` si queremos contrastar la igualdad de las varianzas poblacionales. Con `alternative` indicamos el sentido de la hipótesis alternativa; como ocurría más arriba, `two.sided`, es la opción que se utiliza por defecto y que corresponde al caso de *igual* frente a *distinta*; `greater`, correspondiente al caso de hipótesis alternativa *mayor*, y `less` para el caso de hipótesis alternativa *menor*.

Otro test para analizar la homocedasticidad, especialmente útil cuando tenemos más de dos grupos es el *test de Barlett* aunque, como el anterior, requiere de la normalidad de los datos cuya igualdad de varianzas queremos comparar. Con R se obtiene ejecutando la función `barlett.test`.

Ejemplo 5.1 (continuación)

Para contrastar la igualdad de las varianzas en este ejemplo ejecutamos

```
> var.test(Grupo1, Grupo2, ratio=1)

      F test to compare two variances

data:  Grupo1 and Grupo2
F = 0.3131, num df = 11, denom df = 10, p-value = 0.06976
      (1)
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.08544081 1.10400497
sample estimates:
ratio of variances
 0.3131332
```

El p-valor, obtenido en (1) permite aceptar la hipótesis nula de igualdad de las varianzas aunque no con mucha seguridad.

El test de Barlett lo ejecutamos a continuación

```
> bartlett.test(acido~grupo)
```

```
Bartlett test of homogeneity of variances
```

```
data:  acido by grupo
```

```
Bartlett's K-squared = 3.252, df = 1, p-value = 0.07134
(2)
```

obteniendo en (2) un p-valor que sugiere la aceptación de la homocedasticidad.

5.4. Transformaciones Box-Cox

Como hemos visto, la normalidad y homocedasticidad son dos suposiciones necesarias para poder aplicar tests paramétricos que son los tests más deseados por ser los más potentes.

Una posibilidad a analizar, antes de utilizar tests no paramétricos, es la de si transformando los datos podemos conseguir estas suposiciones, lo que permitiría utilizar tests paramétricos para los datos transformados. Una familia de transformaciones es la *familia Box-Cox*, en donde los datos x eran transformados en $h(x)$ mediante la función

$$h(x) = \begin{cases} \frac{(x+c)^a - 1}{a} & a \neq 0, (x > -c) \\ \log(x+c) & a = 0, (c > 0) \end{cases}$$

en donde a se determina a partir de los datos y c se elige para que sea $x_i + c > 0$, $\forall i = 1, \dots, n$.

Así pues, c será cero si todos los datos son positivos o igual a menos el menor de los datos si algún de ellos es negativo.

La determinación de a y la transformación formal de los datos se pueden hacer con R. La determinación de a se puede hacer con la función `boxcoffit` de la librería `geoR` y la transformación efectiva Box-Cox con la función `bcPower` de la librería `car`. Como siempre, si no dispone en R de alguna de esas librerías las puede obtener de Internet.

Ejemplo 5.2

Los datos que siguen (Afifi y Clark, 1990)

4, 5, 7, 9, 7, 23, 11, 20, 11, 15, 35, 27, 23, 25, 23, 28, 28, 6, 13, 8, 2, 9, 9, 5, 6, 19, 9
9, 8, 45, 9, 2, 5, 2, 19, 4, 19, 8, 5, 7, 11, 7, 5, 4, 7, 7, 4, 6, 7, 15, 23, 28, 5, 2, 15, 9

corresponden a los ingresos de 66 personas encuestadas en Los Ángeles con un nivel de educación de No Graduados. Primero incorporamos estos datos ejecutando

```
> salario<-c(4,5,7,9,7,23,11,20,11,15,35,27,23,25,23,28,28,6,13,8,2,
9,9,5,6,19,9,9,8,45,9,2,5,2,19,4,19,8,5,7,11,7,5,4,7,7,4,6,7,15,23,
28,5,2,15,9,19,20,4,7,9,7,24,9,11,8)
```

Un simple análisis de normalidad sugiere, con el p-valor dado en (1), que los datos no siguen una distribución normal

```
> ks.test(salario,"pnorm",mean(salario),sd(salario))
```

One-sample Kolmogorov-Smirnov test

data: salario

D = 0.2431, p-value = 0.0008195

(1)

alternative hypothesis: two-sided

Para averiguar cuál sería el parámetro a de la transformación de Box-Cox, ejecutamos

```
> library(geoR)
```

```
> boxcofit(salario)
```

Fitted parameters:

```
lambda      beta      sigmasq
0.03745035  2.34983114  0.62697205
```

(2)

El parámetro `lambda`, cuyo valor aparece en (2), resulta igual a $a = 0.03745$. Los datos transformados se obtienen ejecutando (3) y su histograma ejecutando (4), que puede considerarse como el de datos procedentes de una normal.

```
> library(car)
```

```
> trans<-bcPower(salario,0.03745035) (3)
```

```
> hist(trans,prob=T,col=2,main="Histograma de datos transformados") (4)
```

Para confirmarlo ejecutamos de nuevo el test de Kolmogorov-Smirnov, obteniendo ahora un p-valor 0.1748 que admite la normalidad de los datos.

```
> ks.test(trans,"pnorm",mean(trans),sd(trans))
```

One-sample Kolmogorov-Smirnov test

data: trans

D = 0.1359, p-value = 0.1748

alternative hypothesis: two-sided

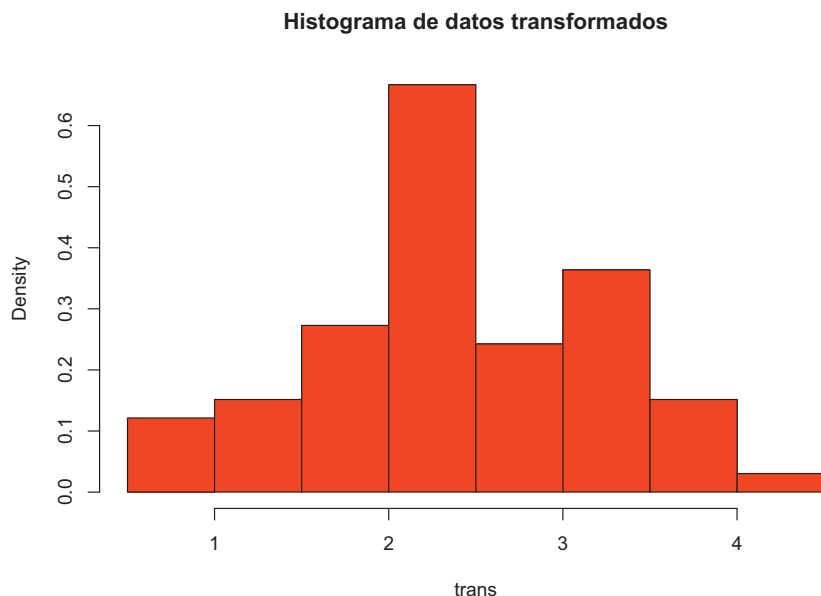


Figura 5.3 : Histograma de los datos transformados

En el caso de una sola población no tiene sentido analizar la homocedasticidad pero conviene resaltar que la transformación Box-Cox consigue, habitualmente, tanto la normalidad como la homocedasticidad de los datos aunque si tenemos más de una población, la elección del parámetro a de la transformación Box-Cox resulta un poco más elaborada.

La utilización de las transformaciones Box-Cox en el análisis de la homocedasticidad está justificada en el caso de que se suponga una correlación entre las medias y las varianzas de cada tratamiento. Es decir si, supuesto que queremos comparar r poblaciones, representamos en un eje de coordenadas los puntos

$$\{(\bar{x}_i, S_i), i = 1, \dots, r\}$$

con S_i la cuasidesviación típica muestral de la población i -ésima, y descubrimos una dependencia que permite ajustar a la nube de puntos de los r pares anteriores, una función de la forma

$$S = c_1 \cdot \bar{x}^\lambda$$

o, equivalentemente, una recta a los logaritmos decimales de ambas

$$\log_{10} S = c_2 + \lambda \log_{10} \bar{x}$$

Transformando ahora los datos con una transformación Box-Cox de $a = 1 - \lambda$ conseguiremos datos con varianza constante.

Ejemplo 5.3

Los datos que aparecen a continuación (Dolkart et al., 1971) muestran las cantidades de albúmina de suero bovino de nitrógeno enlazado producido por tres grupos de ratones diabéticos: los Normales, los Alloxan, y los Alloxan tratados con Insulina.

Normales	156	282	197	297	116	127	119	29	253	122
	349	110	143	64	26	86	122	455	655	14
Alloxan	391	46	469	86	174	133	13	499	168	62
	127	276	176	146	108	276	50	73		
Alloxan+Insulina	82	100	98	150	243	68	228	131	73	18
	20	100	72	133	465	40	46	34	44	

Primero vamos a incorporar los datos ejecutando

```
> Norma<-c(156,282,197,297,116,127,119,29,253,122,349,110,143,64,26,86,122,455,655,14)
> All<-c(391,46,469,86,174,133,13,499,168,62,127,276,176,146,108,276,50,73)
> AllInsu<-c(82,100,98,150,243,68,228,131,73,18,20,100,72,133,465,40,46,34,44)
> ratones<-data.frame(Y=c(Norma,All,AllInsu),Trata=factor(rep(c("Norma","All","AllInsu"),
+ times=c(length(Norma),length(All),length(AllInsu))))))
```

Si utilizáramos para contrastar la normalidad un test de Kolmogorov-Smirnov

```
> ks.test(Norma,"pnorm",mean(Norma),sd(Norma))
One-sample Kolmogorov-Smirnov test
data: Norma
D = 0.2252, p-value = 0.2627
alternative hypothesis: two-sided (1)

> ks.test(All,"pnorm",mean(All),sd(All))
One-sample Kolmogorov-Smirnov test
data: All
D = 0.2383, p-value = 0.2584
alternative hypothesis: two-sided (1)

> ks.test(AllInsu,"pnorm",mean(AllInsu),sd(AllInsu))
One-sample Kolmogorov-Smirnov test
data: AllInsu
D = 0.2327, p-value = 0.2549
alternative hypothesis: two-sided (1)
```

los tres p-valores, marcados con (1) sugieren aceptar la normalidad de los tres conjuntos de datos, pero si simplemente calculamos un histograma del último conjunto de datos,

```
> hist(AllInsu,prob=T)
```

veríamos en la Figura (5.4) una fuerte asimetría a la derecha. Por esta razón es recomendable ejecutar un test de Shapiro-Wilk, seguramente el test más potente en detectar la no normalidad de unos datos.

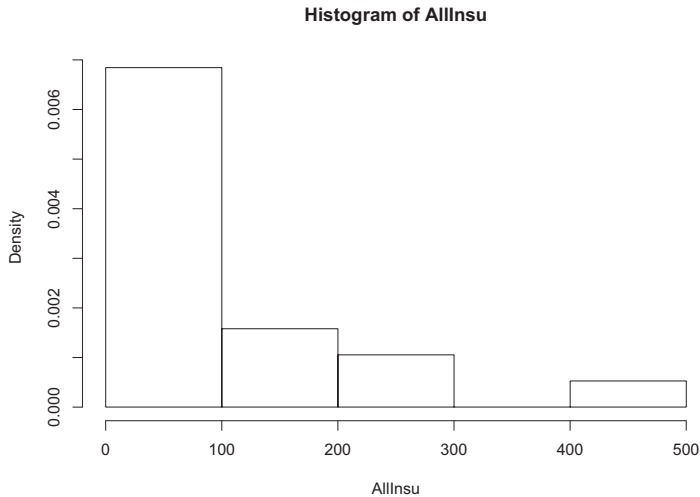


Figura 5.4 : Histograma de AllInsu

Este test se ejecuta a continuación

```
> shapiro.test(Norma)
      Shapiro-Wilk normality test
data:  Norma
W = 0.8433, p-value = 0.004118

> shapiro.test(All)
      Shapiro-Wilk normality test
data:  All
W = 0.8673, p-value = 0.01608

> shapiro.test(AllInsu)
      Shapiro-Wilk normality test
data:  AllInsu
W = 0.7556, p-value = 0.0002771
```

rechazándose la normalidad en los tres casos. Vamos a hacer una transformación Box-Cox siguiendo las indicaciones anteriores. Para ello calculamos primero los logaritmos decimales

de las medias y cuasidesviaciones típicas de los tres conjuntos de datos y el coeficiente de la recta de mínimos cuadrados que se ajusta, dado que existe un fuerte correlación entre las medias y las varianzas de cada tratamiento.

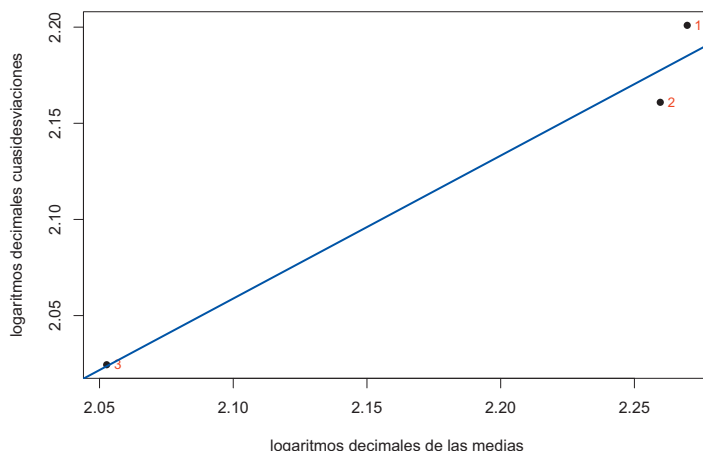


Figura 5.5 : Ajuste para transformación Box-Cox

```
> x1<-c(log10(mean(Norma)),log10(sd(Norma)))
> x2<-c(log10(mean(All)),log10(sd(All)))
> x3<-c(log10(mean(AllInsu)),log10(sd(AllInsu)))

> X<-c(x1[1:1],x2[1:1],x3[1:1])
> Y<-c(x1[2:2],x2[2:2],x3[2:2])

> plot(X,Y,xlab="logaritmos decimales de las medias",
+ ylab="logaritmos decimales cuasidesviaciones",pch=16)
> text(X,Y,adj=-1,cex=0.8,col=2)
> recta<-lm(Y~X)
> abline(recta,col=4,lwd=2)
> cor(X,Y)
[1] 0.9843958

> recta
Call:
lm(formula = Y ~ X)
Coefficients:
(Intercept)          X
      0.4975       0.7435
```

El parámetro a de la transformación Box-Cox

$$h(x) = \frac{(x+c)^a - 1}{a}$$

será, por tanto, $a = 1 - 0'7435 = 0'2565$. Dado que todas las observaciones son positivas, será $c = 0$, con lo que los datos deben de transformarse por la fórmula

$$h(x) = \frac{x^{0'2565} - 1}{0'2565}$$

```
> ratonestrans<-data.frame((((ratones[,1])^0.2565)-1)/0.2565, ratones[,2])

> Normatrans<-ratonestrans[1:20,1]
> Alltrans<-ratonestrans[21:38,1]
> AllInsutrans<-ratonestrans[39:57,1]

> shapiro.test(Normatrans)
      Shapiro-Wilk normality test
data:  Normatrans
W = 0.9736, p-value = 0.8288

> shapiro.test(Alltrans)
      Shapiro-Wilk normality test
data:  Alltrans
W = 0.9763, p-value = 0.9037

> shapiro.test(AllInsutrans)
      Shapiro-Wilk normality test
data:  AllInsutrans
W = 0.963, p-value = 0.6333
```

La normalidad puede admitirse ahora. La homocedasticidad la contrastamos con el test de Bartlett

```
> bartlett.test(ratonestrans[,1]~ratonestrans[,2],data=ratonestrans)
      Bartlett test of homogeneity of variances
data:  ratonestrans[, 1] by ratonestrans[, 2]
Bartlett's K-squared = 0.709, df = 2, p-value = 0.7015
                                     (2)
```

El p-valor, marcado con (2), indica que se puede aceptar ésta.

5.5. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones normales independientes

La situación considerada en esta sección es la de datos procedentes de dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$, con tamaños muestrales n_1 y n_2 respectivamente, representando $\bar{x}_1$, S_1^2 y $\bar{x}_2$, S_2^2 la media y cuasivarianza de la primera y segunda muestra respectivamente.

$$\begin{array}{l} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{array}$$

σ_1 y σ_2 **conocidas**

En este caso el test óptimo es

$$\begin{array}{l} \bullet \text{ Se acepta } H_0 \text{ si } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \\ \bullet \text{ Se rechaza } H_0 \text{ si } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2} \end{array}$$

σ_1 y σ_2 **desconocidas. Muestras pequeñas**

Aquí habrá que distinguir los casos en que las varianzas poblacionales puedan considerarse iguales y aquellos en los que no puedan ser consideradas iguales.

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales se pueden considerar iguales, entonces el test óptimo es

• Se acepta H_0 si • Se rechaza H_0 si	$\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \leq t_{n_1+n_2-2; \alpha/2}$ $\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} > t_{n_1+n_2-2; \alpha/2}$
---	--

(b) $\sigma_1 \neq \sigma_2$

En el caso de que las varianzas poblacionales no puedan considerarse iguales, el test óptimo, denominado test de Welch, es

• Se acepta H_0 si • Se rechaza H_0 si	$\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq t_{f; \alpha/2}$ $\frac{ \bar{x}_1 - \bar{x}_2 }{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > t_{f; \alpha/2}$
---	--

en donde f son los grados de libertad dados por la aproximación de Welch estudiados en el Capítulo 3.

Ejemplo 5.4

En un artículo del 27 de Mayo de 2001 del diario “The Arizona Republic”, aparecen los datos de las concentraciones de arsénico en partes por billón (americano) en el agua potable de 10 barrios de Phoenix, capital del estado norteamericano de Arizona (columna izquierda de la tabla), y de 10 zonas rurales de dicho estado (columna derecha de la tabla). Los datos fueron los siguientes:

Phoenix Centro	3	Rimrock	48
Chandler	7	Goodyear	44
Gilbert	25	New River	40
Glendale	10	Apache Junction	38
Mesa	15	Buckeye	33
Paradise Valley	6	Nogales	21
Peoria	12	Black Canyon City	20
Scottsdale	25	Sedona	12
Sun City	7	Casa Grande	18
Tempe	15	Payson	1

Suponiendo que los dos grupos de datos proceden de poblaciones normales, para analizar si existen diferencias significativas entre ellos debemos analizar primero si las varianzas pueden considerarse como iguales o distintas. Para ello, comenzaremos incluyendo los datos y luego contrastando la igualdad de las varianzas poblacionales,

```
> ciudad<-c(3,7,25,10,15,6,12,25,7,15)
> campo<-c(48,44,40,38,33,21,20,12,18,1)

> var.test(ciudad,campo)
```

F test to compare two variances

```
data: ciudad and campo
F = 0.2473, num df = 9, denom df = 9, p-value = 0.04936
(1)
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.06143758 0.99581888
sample estimates:
ratio of variances
 0.2473473
```

El p-valor obtenido en (1) no es nada concluyente. Si suponemos que las varianzas son iguales, el test sobre la hipótesis nula de igualdad de ambos grupos de datos, es decir, la hipótesis nula $H_0 : \mu_1 = \mu_2$ frente a la alternativa $H_1 : \mu_1 \neq \mu_2$ se resuelve ejecutando (2)

```
> t.test(ciudad,campo,var.equal=T) (2)
```

Two Sample t-test

```
data: ciudad and campo
t = -2.7669, df = 18, p-value = 0.01270
(3)
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-26.389418 -3.610582
sample estimates:
mean of x mean of y
 12.5      27.5
```

El p-valor 0.0127 obtenido en (3) y sugiere el rechazo de H_0 .

Por tanto, existe suficiente evidencia para concluir que la concentración media de arsénico en el agua potable en las zonas rurales de Arizona es distinta que en su capital Phoenix.

Ejemplo 5.5

Los siguientes datos corresponden a un estudio sobre trombosis (van Oost et al., 1983) en donde se midió la cantidad de tromboglobulina urinaria eliminada por 12 pacientes normales y 12 pacientes con diabetes.

Normales:

4'1 , 6'3 , 7'8 , 8'5 , 8'9 , 10'4 , 11'5 , 12 , 13'8 , 17'6 , 24'3 , 37'2

Diabéticos:

11'5 , 12'1 , 16'1 , 17'8 , 24 , 28'8 , 33'9 , 40'7 , 51'3 , 56'2 , 61'7 , 69'2

Supuesto que ambos grupos de datos proceden de distribuciones normales, ¿puede aceptarse la igualdad de las medias de ambas poblaciones a nivel 0'05?

Se trata de la comparación de medias de dos poblaciones normales independientes y muestras pequeñas, siendo las varianzas poblacionales desconocidas, para lo que necesitamos primero analizar si éstas pueden considerarse iguales. Para ello contrastamos la hipótesis nula $H_0 : \sigma_1^2 = \sigma_2^2$ frente a la $H_0 : \sigma_1^2 \neq \sigma_2^2$. Para ello, primero incorporamos los datos y luego ejecutamos el test anterior,

```
> normal<-c(4.1,6.3,7.8,8.5,8.9,10.4,11.5,12,13.8,17.6,24.3,37.2)
> diabetico<-c(11.5,12.1,16.1,17.8,24,28.8,33.9,40.7,51.3,56.2,61.7,69.2)

> var.test(normal,diabetico)
```

F test to compare two variances

data: normal and diabetico

F = 0.2058, num df = 11, denom df = 11, p-value = 0.01435

(1)

alternative hypothesis: true ratio of variances is not equal to 1

95 percent confidence interval:

0.05923198 0.71472776

sample estimates:

ratio of variances

0.2057541

El p-valor obtenido en (1) sugiere rechazar la igualdad de las varianzas por lo que contrastaremos la hipótesis nula de igualdad de las medias de ambos grupos, $H_0 : \mu_1 = \mu_2$ en el caso de poblaciones normales, muestras pequeñas y varianzas desconocidas y distintas, es decir, mediante el test de Welch ejecutando

```
> t.test(normal,diabetico,var.equal=F)
Welch Two Sample t-test
```

```

data: normal and diabetico
t = -3.3838, df = 15.343, p-value = 0.003982
                                (2)
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -35.41024 -8.07309
sample estimates:
mean of x mean of y
 13.53333 35.27500

```

El p-valor dado en (2) sugiere rechazar la hipótesis nula de igualdad de ambos grupos de datos.

En el caso de que se desee contrastar la hipótesis unilateral, las fórmulas serían las siguientes, en donde sólo hemos considerado un sentido de unilateralidad. Intercambiando los papeles de las dos poblaciones tendríamos las análogas.

Como en el apartado anterior, habrá que distinguir si las varianzas poblacionales pueden considerarse conocidas o no, y en ese caso, si pueden admitirse como iguales.

$$\begin{array}{l}
 H_0 : \mu_1 \geq \mu_2 \\
 H_1 : \mu_1 < \mu_2
 \end{array}$$

σ_1 y σ_2 **conocidas**

En este caso el test óptimo es

$$\begin{array}{l}
 \bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq z_{1-\alpha} \\
 \bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{1-\alpha}
 \end{array}$$

σ_1 y σ_2 **desconocidas. Muestras pequeñas**

(a) $\sigma_1 = \sigma_2$

Si las varianzas poblacionales pueden suponerse iguales y las muestras no tienen ambas, tamaños suficientemente grandes, el test óptimo es

• Se acepta H_0 si	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \geq t_{n_1+n_2-2;1-\alpha}$
• Se rechaza H_0 si	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < t_{n_1+n_2-2;1-\alpha}$

(b) $\sigma_1 \neq \sigma_2$

Si las varianzas poblacionales son distintas, el test óptimo es

• Se acepta H_0 si	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \geq t_{f;1-\alpha}$
• Se rechaza H_0 si	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} < t_{f;1-\alpha}$

Ejemplo 5.1 (continuación)

Si queremos analizar si el aporte de vitamina C en ambientes muy fríos disminuye los niveles de ácido ascórbico en sangre, las hipótesis a contrastar serán $H_0 : \mu_1 \geq \mu_2$ frente a $H_1 : \mu_1 < \mu_2$.

Ya analizamos que los niveles de ácido ascórbico siguen distribuciones normales en ambas poblaciones así como que se puede admitir la igualdad de las varianzas.

Para ejecutar el test propuesto ejecutaremos

```
> t.test(Grupo1,Grupo2,alternative="less",var.equal=T)
```

Two Sample t-test

data: Grupo1 and Grupo2

t = -2.6989, df = 21, p-value = 0.006722

(1)

alternative hypothesis: true difference in means is less than 0

95 percent confidence interval:


```

-Inf -1.722055
sample estimates:
mean of x mean of y
14.86667 19.61818

```

Un p-valor tan pequeño, obtenido en (1), sugiere rechazar H_0 e inferir, en base a estos datos, que la administración de vitamina C en ambientes muy fríos disminuye los niveles de ácido ascórbico en la sangre.

5.6. Contraste de hipótesis relativas a la diferencia de medias de dos poblaciones independientes no necesariamente normales. Muestras grandes

La situación que se estudia en esta sección es la de dos muestras independientes $X_1, \dots, X_{n_1}$ e $Y_1, \dots, Y_{n_2}$, de tamaños similares y suficientemente grandes ($n_1 + n_2 > 30$).

Precisamente por esta razón no se requiere normalidad en las distribuciones modelo.

$$\begin{aligned}
 H_0 : \mu_1 &= \mu_2 \\
 H_1 : \mu_1 &\neq \mu_2
 \end{aligned}$$

σ_1 y σ_2 **conocidas**

En este caso el test óptimo es

- Se acepta H_0 si $\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}$
- Se rechaza H_0 si $\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$

σ_1 y σ_2 **desconocidas**

Si las varianzas poblacionales no se suponen conocidas —situación por otro lado habitual—, el test óptimo es

$$\begin{array}{l}
 \bullet \text{ Se acepta } H_0 \text{ si } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq z_{\alpha/2} \\
 \bullet \text{ Se rechaza } H_0 \text{ si } \frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > z_{\alpha/2}
 \end{array}$$

Ejemplo 5.6

Se realizó un estudio a lo largo de 12 meses, en el cual se recogieron datos sobre las mujeres que daban a luz en hospitales de Tasmania, sobre del uso de *Syntocinon*, un medicamento utilizado para provocar el parto.

El grupo 1 fue un grupo control formado por mujeres que no usaron el medicamento, y el grupo 2 el formado por mujeres que lo usaron dentro de un periodo de dos horas desde que rompieron aguas.

Los datos, en horas, desde que rompieron aguas hasta el momento del parto fueron (Fuente: Profess. J. Correy, Depart. of Obstets., University of Tasmania)

$$\begin{array}{lll}
 n_1 = 315 & \bar{x}_1 = 9'43 & S_1^2 = 32'4616 \\
 n_2 = 301 & \bar{x}_2 = 9'14 & S_2^2 = 26'2455
 \end{array}$$

A nivel $\alpha = 0'05$, ¿puede inferirse una diferencia significativa entre ambos grupos?

Como es

$$\frac{|\bar{x}_1 - \bar{x}_2|}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{|9'43 - 9'14|}{\sqrt{\frac{32'4616}{315} + \frac{26'2455}{301}}} = 0'6649 < 1'96 = z_{0'025}$$

se acepta la no existencia de diferencias significativas entre ambos grupos, es decir, se acepta la hipótesis $H_0 : \mu_1 = \mu_2$.

$$\begin{array}{l}
 H_0 : \mu_1 \leq \mu_2 \\
 H_1 : \mu_1 > \mu_2
 \end{array}$$

σ_1 y σ_2 conocidas

Si las varianzas de las poblaciones son, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_\alpha \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_\alpha
 \end{aligned}$$

σ_1 y σ_2 desconocidas

Caso de que se desconozcan las varianzas de las poblaciones, el test óptimo es

$$\begin{aligned}
 &\bullet \text{ Se acepta } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \leq z_\alpha \\
 &\bullet \text{ Se rechaza } H_0 \text{ si } \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > z_\alpha
 \end{aligned}$$

Ejemplo 5.7

Los siguientes datos proceden de un estudio del Western Collaborative Group llevado a cabo en California en 1960-1961. En concreto corresponde a 40 individuos de ese estudio de peso elevado, con los que se formaron dos grupos: El Grupo A formado por 20 individuos estresados, ambiciosos y agresivos, y el Grupo B formado por 20 individuos relajados, no competitivos y no estresados. Se midieron en ambos grupos los niveles de colesterol en mgr. por 100 ml. obteniéndose los siguientes datos:

Grupo A:

233 , 291 , 312 , 250 , 246 , 197 , 268 , 224 , 239 , 239
254 , 276 , 234 , 181 , 248 , 252 , 202 , 218 , 212 , 325

Grupo B:

344 , 185 , 263 , 246 , 224 , 212 , 188 , 250 , 148 , 169
226 , 175 , 242 , 252 , 153 , 183 , 137 , 202 , 194 , 213

¿Existen diferencias significativas a favor de alguno de los dos grupos?

La pregunta se refiere a inferencias sobre las medias de dos poblaciones independientes y, al ser los tamaños muestrales suficientemente grandes y semejantes, no necesitamos la normalidad de las poblaciones de donde proceden los datos.

Aunque no estaría mal del todo analizar simplemente si existen diferencias significativas entre ambos grupos contrastando la hipótesis nula de ser las medias de ambas poblaciones iguales, $H_0 : \mu_1 = \mu_2$, dado que, como veremos un poco más abajo, es $\bar{x}_1 = 245'05$ y $\bar{x}_2 = 210'3$, la hipótesis de interés es analizar si esa diferencia entre ambas medias muestrales implica una diferencia significativa entre las medias poblacionales, es decir, resulta de interés contrastar la hipótesis $\mu_1 > \mu_2$ por lo que, siguiendo la metodología propia de los tests de hipótesis ésta debería de ser la hipótesis alternativa, y deberíamos contrastar $H_0 : \mu_1 \leq \mu_2$ frente a $H_1 : \mu_1 > \mu_2$ en el caso que nos ocupa de ser las varianzas poblacionales desconocidas, rechazando la hipótesis nula si

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} > z_\alpha.$$

Con R fácilmente se obtiene que

```
> x1<-c(233,291,312,250,246,197,268,224,239,239,254,276,234,181,248,252,202,218,212,325)
> x2<-c(344,185,263,246,224,212,188,250,148,169,226,175,242,252,153,183,137,202,194,213)
> mean(x1)
[1] 245.05
> mean(x2)
[1] 210.3
> var(x1)
[1] 1342.366
> var(x2)
[1] 2336.747
```

con lo que será

$$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} = \frac{245'05 - 210'30}{\sqrt{\frac{1342'37}{20} + \frac{2336'75}{20}}} = 2'56.$$

Como en todo test de hipótesis, la mejor forma de obtener una conclusión es mediante el cálculo del p-valor. Es decir, del cálculo de

$$P\{Z > 2'56\} = 0'0052$$

obtenido al ejecutar

```
> 1-pnorm(2.56)
[1] 0.005233608
```

Un p-valor tan pequeño indica rechazar claramente la hipótesis nula y concluir que puede admitirse un nivel medio de colesterol significativamente mayor en el Grupo A.

Si utilizamos R, el test de hipótesis deberíamos hacerlo ejecutando

```
> t.test(x1,x2,alternative="greater")

Welch Two Sample t-test

data:  x1 and x2
t = 2.5621, df = 35.413, p-value = 0.007405
(1)
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 11.84155      Inf
sample estimates:
mean of x mean of y
 245.05    210.30
```

obteniendo en (1) de nuevo un p-valor muy pequeño. La pequeña diferencia que se obtiene con el cálculo anterior se debe a que los cálculos de R se hacen con la t de Student, la cual sólo converge a la normal (la que utilizamos en los cálculos de más arriba) cuando el tamaño muestral es muy grande. No obstante, las conclusiones son muy claras.

5.7. El contraste de Wilcoxon-Mann-Whitney

En las secciones anteriores estudiamos contrastes paramétricos para comparar dos poblaciones cuando, o bien se puede admitir que los datos proceden de poblaciones normales o bien los tamaños muestrales son suficientemente grandes. Si no estamos en ninguna de estas dos situaciones, debemos utilizar tests no paramétricos, como el que analizamos aquí, en el que la hipótesis nula de igualdad de las dos poblaciones independientes a comparar se expresa mediante sus medianas poblacionales, M_X y M_Y . Además, este test requiere que los datos sean de tipo continuo.

$H_0 : M_X = M_Y$ $H_1 : M_X \neq M_Y$
--

La idea de este contraste consiste en medir las magnitudes de los valores de la segunda muestra (de tamaño n) en relación con los de la primera (de tamaño m), es decir, las posiciones de la segunda muestra en la muestra conjunta de las dos. Si observamos que la mayoría de estos valores de la segunda muestra están hacia el principio o hacia el final de la muestra conjunta, deberemos rechazar la hipótesis nula de igualdad de ambas poblaciones.

En concreto, si llamamos U al estadístico de contraste que mide el número de datos de la segunda muestra que preceden estrictamente a cada uno de los de la primera muestra, valores muy grandes o muy pequeños de U desacreditarán

la hipótesis nula de igualdad de ambas poblaciones. Así pues, fijado un nivel de significación α ,

- Se acepta H_0 si $m \cdot n - u_{m,n;\alpha/2} < U < u_{m,n;\alpha/2}$
- Se rechaza H_0 si $U \leq m \cdot n - u_{m,n;\alpha/2}$ ó $U \geq u_{m,n;\alpha/2}$

en donde $u_{m,n;\alpha/2}$ es el menor número entero tal que

$$P\{U \geq u_{m,n;\alpha/2}\} \leq \frac{\alpha}{2}.$$

Para ejecutar este test con R, utilizaremos de nuevo la función antes introducida,

```
wilcox.test(x,y,alternative="two.sided",mu=0)
```

en donde incluiremos en el primer argumento x el vector de observaciones de una de las dos poblaciones a comparar y en el segundo, y , los datos de la otra población. El resto de los argumentos son los anteriormente explicados.

Ejemplo 5.8

Se realizó un estudio con objeto de averiguar si el número de pulsaciones por minuto puede considerarse igual entre los hombres y mujeres de una determinada población.

Para ello se eligieron al azar 12 hombres y 12 mujeres de la mencionada población obteniéndose los siguientes datos

Individuo	1	2	3	4	5	6	7	8	9	10	11	12
<i>Hombres</i>	74	77	71	76	79	74	83	79	83	72	79	77
<i>Mujeres</i>	81	84	80	73	78	80	82	84	80	84	75	82

Si representamos por X la pulsación en la población de hombres y por Y la pulsación en la de mujeres, las hipótesis que se quieren contrastar son $H_0 : M_X = M_Y$, frente a $H_1 : M_X \neq M_Y$.

Para este ejemplo, incorporamos los datos en (1) y (2) y ejecutamos la función en (3). No hemos incluido los argumentos `alternative` ni `mu` porque vamos a ejecutar los que toma por defecto, respectivamente, la igualdad de las medianas de ambas poblaciones y que su diferencia es 0.

```
> x<-c(74,77,71,76,79,74,83,79,83,72,79,77) (1)
```

```
> y<-c(81,84,80,73,78,80,82,84,80,84,75,82) (2)
```

```
> wilcox.test(x,y) (3)
```

Wilcoxon rank sum test with continuity correction

```
data: x and y
W = 35, p-value = 0.03446
(4)
alternative hypothesis: true location shift is not equal to 0
```

Los resultados del estadístico de contraste, 35, y de su p-valor, 0'03446, aparecen en (4). Este p-valor no es concluyente, pero indica rechazar la hipótesis nula de igualdad entre las medianas de ambas poblaciones a un nivel de significación $\alpha = 0'05$ por ser este valor, mayor que el p-valor lo que indica que el estadístico toma un valor perteneciente a la región crítica del test.

De nuevo, en la hipótesis unilaterales sólo consideraremos una de ellas.

$H_0 : M_X \leq M_Y$ $H_1 : M_X > M_Y$
--

Fijado un nivel de significación α

- | |
|---|
| <ul style="list-style-type: none"> • Se acepta H_0 si $U < u_{m,n;\alpha}$ • Se rechaza H_0 si $U \geq u_{m,n;\alpha}$ |
|---|

en donde $u_{m,n;\alpha}$ es el menor número entero tal que

$$P\{U \geq u_{m,n;\alpha}\} \leq \alpha.$$

Las hipótesis H_0 y H_1 las hemos expresado en función de las medianas poblacionales, queriendo destacar con ello el hecho de que si se acepta, por ejemplo, la hipótesis alternativa, $H_1 : M_X > M_Y$, se concluye con que la variable en observación tiende a tomar valores significativamente mayores en la población denominada X que en la población denominada Y .

5.8. Análisis de la Varianza

En las secciones anteriores hemos considerado el caso de comparación de dos poblaciones. Si el número de grupos a comparar es tres o más de tres, deberemos utilizar las técnicas estudiadas en estas últimas secciones. Por ejemplo, si tenemos r grupos a comparar, nuestros datos estarán en una tabla como la siguiente

Tratamiento	Observaciones			
1	x_{11}	x_{12}	$\cdots$	x_{1n_1}
2	x_{21}	x_{22}	$\cdots$	x_{2n_2}
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
r	x_{r1}	x_{r2}	$\cdots$	x_{rn_r}

En esta sección estudiaremos el *Análisis de la Varianza*, que permite contrastar la hipótesis nula de igualdad de los efectos medios de las r poblaciones o grupos de datos $H_0 : \mu_1 = \mu_2 = \dots = \mu_r$ frente a la alternativa de no ser iguales todos estos efectos medios, $H_1 : \text{no todos son iguales}$, utilizando $n_1, \dots, n_r$ individuos tomados al azar de cada una de las r poblaciones a comparar, siendo $n = n_1 + \dots + n_r$ el número total de individuos de la muestra.

Las suposiciones que esta técnica requiere son, básicamente, que los datos sean de tipo continuo con distribución normal en cada grupo de datos a comparar y que tengan la misma varianza los r grupos de datos (suposición de *homocedasticidad*). El análisis de ambas suposiciones ya lo hemos abordado en secciones anteriores.

La idea del Análisis de la Varianza es descomponer la variación existente en los datos en dos fuentes de variación: una, la debida a las poblaciones a comparar, aquí denominados *Tratamientos*, y otra, la debida al azar. Si la primera fuente de variación, designada por SST_i es grande en comparación con la otra, denotada por SSE , rechazaremos la hipótesis nula de igualdad de los efectos medios de las poblaciones o grupos de datos a comparar. Por esta razón, en esencia, el estadístico de contraste será el cociente de ambas fuentes de variación SST_i/SSE , aunque hay que *estandarizarlas* para que el cociente tenga una distribución conocida (una F de Snedecor) y poder medir así sus variaciones en términos de probabilidades.

Los cálculos se presentan en una tabla denominada ANOVA, que es lo que nos da el ordenador en donde aparece el valor del estadístico de contraste

$$F = \frac{SST_i/(r-1)}{SSE/(n-r)}$$

que seguirá una distribución F de Snedecor con $(r-1, n-r)$ grados de libertad.

F. de variación	Suma de cuadrados	g.l.	c. medios	Estadístico
Tratamientos	$SST_i = \sum_{i=1}^r \frac{T_i^2}{n_i} - \frac{T^2}{n}$	$r - 1$	$\frac{SST_i}{r - 1}$	$\frac{SST_i/(r - 1)}{SSE/(n - r)}$
Residual	$SSE = SST - SST_i$	$n - r$	$\frac{SSE}{n - r}$	
Total	$SST = \sum_{i=1}^r \sum_{j=1}^{n_i} x_{ij}^2 - \frac{T^2}{n}$	$n - 1$		

Contraste de hipótesis

Si $F_{r-1, n-r; \alpha}$ es, como siempre, el valor de la abscisa de una F de Snedecor con $(r - 1, n - r)$ grados de libertad que deja a la derecha un área de probabilidad α ,

- Se acepta H_0 si $F < F_{r-1, n-r; \alpha}$
- Se rechaza H_0 si $F \geq F_{r-1, n-r; \alpha}$

Teniendo perfecto sentido, al ser éste un contraste de hipótesis, el cálculo e interpretación del p-valor del test.

Análisis de la Varianza con R

La función de R que vamos a utilizar para ejecutar el Análisis de la Varianza es

```
aov(modelo, datos)
```

incluyendo en el argumento `modelo` la variable dependiente cuantitativa observada, en función del factor que define las poblaciones a comparar. En `datos` incluiremos las observaciones que tendrán que venir expresadas en formato *data frame*.

Ejemplo 5.9

Con objeto de analizar si existen diferencias en el aumento de peso entre tres dietas, se decidió someter a 5 ratones a cada una de ellas, obteniéndose los siguientes aumentos de peso

Dieta	Aumento de peso					T_i	$\bar{x}_i$
A	32	37	34	33	30	166	33'2
B	36	38	37	30	34	175	35
C	35	30	36	29	31	161	32'2
						502	

Supuesto que hemos verificado las suposiciones de normalidad y homocedasticidad, para contrastar $H_0 : \mu_A = \mu_B = \mu_C$ frente a la alternativa de no ser iguales todos estos efectos medios, $H_1 : \text{alguna distinta}$, primero creamos los datos, los cuales tendrán que venir en formato *data frame* para que los entienda R, mediante la secuencia (1), (2) y (3),

```
> peso<-c(32,37,34,33,30,36,38,37,30,34,35,30,36,29,31) (1)
> dieta<-c("A","A","A","A","A","B","B","B","B","C","C","C","C","C") (2)
> ejemplo<-data.frame(dieta,peso) (3)
```

Para obtener la tabla de Análisis de la Varianza ejecutamos (4) y (5)

```
> resul<-aov(peso~dieta,ejemplo) (4)
> summary(resul) (5)
```

```
> summary(resul)
          Df Sum Sq Mean Sq F value Pr(>F)
dieta      2  20.13   10.07    1.144  0.351
          (6)
```

```
Residuals  12 105.60    8.80
```

El p-valor del test, que aparece en (6) indica, claramente, la aceptación de la hipótesis nula de igualdad de los efectos medias de las tres dietas.

5.8.1. Comparaciones Múltiples

En el ejemplo anterior hemos aceptado la hipótesis nula de igualdad de los efectos medios de las poblaciones a comparar pero, en muchas ocasiones, rechazaremos esta hipótesis, pudiendo hacer *Comparaciones Múltiples* entre los diversos tratamientos sobre los que hemos rechazado la igualdad común de todos ellos, con la idea de formar grupos de tratamientos equivalentes.

La primera idea que se le ocurrirá al lector es la de hacer tests de comparación de dos poblaciones, de nivel α , formando grupos de dos tratamientos. Este método es erróneo porque, en ese caso, el nivel de significación global ya no sería α . En este apartado expondremos tests que sí tienen en cuenta este problema, tests que se denominan de *comparaciones múltiples*.

Contraste de Tukey HSD

Este contraste se basa en calcular el valor HSD , definido por

$$HSD = q_{r,n-r;\alpha} \sqrt{\frac{SSE/(n-r)}{n/r}}$$

y declarar significativa cualquier diferencia que exceda dicho valor.

En este test se requiere que el tamaño muestral de cada tratamiento sea el mismo.

Con R haremos comparaciones múltiples utilizando la función

`TukeyHSD(x, conf.level=0.95)`

cuyo primer argumento `x` debe ser un objeto creado con la función `aov`. El segundo es el `1 -` el nivel de significación (coeficiente de confianza del intervalo de confianza/región de aceptación) de los tests donde la hipótesis nula es la igualdad de las medias de las poblaciones comparadas.

Ejemplo 5.10

En un estudio sobre el efecto de la glucosa en la eliminación de insulina, fueron tratados especímenes de tejidos pancreáticos de animales experimentales con cinco estimulantes diferentes. Más tarde fue determinada la cantidad de insulina eliminada obteniéndose los siguientes resultados:

Estimulante	Observaciones							
1	1'53	1'61	3'75	2'89	3'26	2'83	2'86	2'59
2	3'15	3'96	3'59	1'89	1'45	3'49	1'56	2'44
3	3'89	4'80	3'68	5'70	5'62	5'79	4'75	5'33
4	8'18	5'64	7'36	5'33	8'82	5'26	8'75	7'10
5	5'86	5'46	5'69	6'49	7'81	9'03	7'49	8'98

Se quiere saber si existe diferencia entre los estimulantes en relación con la cantidad de insulina eliminada. Es decir, se trata de contrastar la hipótesis $H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ frente a $H_1 : \text{alguna distinta}$, utilizando un diseño completamente aleatorizado.

Para resolver esta problema con R, primero incorporamos los datos a partir de (1), ejecutamos el Análisis de la Varianza en (2) obteniendo la tabla ANOVA con (3). En (4) se observa un p-valor casi cero lo que lleva a rechazar la igualdad de los efectos medios de los cinco estimulantes. El contraste HSD de Tukey, a nivel 0'05, se obtiene ahora ejecutando (5)

```
> insulina<-c(1.53,1.61,3.75,2.89,3.26,2.83,2.86,2.59,3.15,3.96,3.59,      (1)
+ 1.89,1.45,3.49,1.56,2.44,3.89,4.8,3.68,5.7,5.62,5.79,4.75,5.33,8.18,
+ 5.64,7.36,5.33,8.82,5.26,8.75,7.1,5.86,5.46,5.69,6.49,7.81,9.03,7.49,8.98)
> estimula<-factor(rep(LETTERS[1:5],c(8,8,8,8,8)))
> ejemplo2<-data.frame(estimula,insulina)

> resul2<-aov(insulina~estimula,ejemplo2)                                (2)
> summary(resul2)                                                         (3)
      Df Sum Sq Mean Sq F value    Pr(>F)
estimula    4 154.920   38.730   29.755 7.956e-11 ***
```

```

(4)
Residuals    35  45.557    1.302

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> TukeyHSD(resul2)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = insulina ~ estimula, data = ejemplo2)

$estimula
      diff      lwr      upr      p adj
B-A 0.02625 -1.6138197 1.66632 0.9999989
C-A 2.28000  0.6399303 3.92007 0.0027393
D-A 4.39000  2.7499303 6.03007 0.0000000
E-A 4.43625  2.7961803 6.07632 0.0000000
C-B 2.25375  0.6136803 3.89382 0.0031151
D-B 4.36375  2.7236803 6.00382 0.0000001
E-B 4.41000  2.7699303 6.05007 0.0000000
D-C 2.11000  0.4699303 3.75007 0.0062262
E-C 2.15625  0.5161803 3.79632 0.0049938
E-D 0.04625 -1.5938197 1.68632 0.9999897

```

Los intervalos (regiones de aceptación) obtenidos a partir de (6), cuyo extremo inferior está encabezado con **lwr** y el superior con **upr**, que contengan al cero implicarán la igualdad de los efectos medios cuyas letras aparecen al comienzo de la línea. Así, por ejemplo, el primer intervalo de aceptación es $[-1'61, 1'66]$ el cual, al contener al cero, implica la igualdad de los efectos medios de los tratamientos B-A. De esta manera vemos que podemos considerar tres clases de tratamientos equivalentes: el $\{A, B\}$, $\{C\}$, $\{D, E\}$. La última columna nos da los p-valores de los tests, los cuales confirman que se obtienen tres grupos de tratamientos “equivalentes”, el $\{A, B\}$, el $\{C\}$ y el $\{D, E\}$.

Hemos puesto entre comillas lo de equivalentes, porque las clasificaciones proporcionadas por los tests de comparaciones múltiples no tiene porqué ser disjuntas. Es decir, puede darse el caso de no existir diferencias significativas entre, por ejemplo, el primer y segundo tratamiento, no existir diferencias significativas entre el segundo y el tercero, y sí existir diferencias significativas entre el primero y el tercero.

Varianzas distintas: Test de Welch

R tiene una función que puede utilizarse cuando no puede admitirse la igualdad de la varianzas, la cual ejecuta un test similar a la aproximación de Welch en la comparación de dos poblaciones independientes. Se trata de la función `oneway.test`.

Ejemplo 5.9 (continuación)

Si para los datos del ejemplo 5.9 no se hubiera podido aceptar la igualdad de las varianzas o ésta fuera dudosa, ejecutaríamos (1) obteniendo en (2) un p-valor, de nuevo lo suficientemente alto como para aceptar la hipótesis nula de igualdad de los efectos medios de las tres

dietas.

```
> oneway.test(peso~dieta,ejemplo)
```

(1)

One-way analysis of means (not assuming equal variances)

data: peso and dieta

F = 0.9462, num df = 2.000, denom df = 7.927, p-value = 0.428

(2)

5.9. Contraste de Kruskal-Wallis

Este contraste utiliza los rangos de las observaciones es decir, sus ordenaciones en cada grupo que se pueden expresar en la forma:

Tratamiento	Rangos de las observaciones		Sumas de los rangos
1	r_1	$\cdots r_{n_1}$	$R_1 = \sum_{i=1}^{n_1} r_i$
2	r_{n_1+1}	$\cdots r_{n_1+n_2}$	$R_2 = \sum_{i=1}^{n_2} r_{n_1+i}$
$\vdots$	$\vdots$	$\cdots \vdots$	$\vdots$
r	$r_{n_1+\cdots+n_{r-1}+1}$	$\cdots r_n$	$R_r = \sum_{i=1}^{n_r} r_{n_1+\cdots+n_{r-1}+i}$

y está basado en el hecho de que, si es cierta la hipótesis nula de igualdad de los efectos medios de los r tratamientos, no debería existir tendencia en la suma de los rangos de cada tratamiento, R_i . El estadístico

$$T = \left[\frac{12}{n(n+1)} \sum_{i=1}^r \frac{R_i^2}{n_i} \right] - 3(n+1)$$

recoge esta idea, rechazándose H_0 cuando T tome valores significativamente grandes.

Contraste de hipótesis

Así pues, fijado un nivel de significación α , se define el siguiente contraste

- Se acepta H_0 si $T < t_\alpha$
- Se rechaza H_0 si $T \geq t_\alpha$

en donde por t_α representamos el valor de la abscisa de la distribución de T que deja a la derecha una área de probabilidad α ,

$$P\{T \geq t_\alpha\} = \alpha.$$

La distribución de T es complicada pero con R se puede ejecutar este test fácilmente. La función de R que utilizaremos para ejecutarlo es

`kruskal.test(modelo,datos)`

incluyendo, como más arriba, en el argumento `modelo` la variable dependiente cuantitativa observada, en función del factor que define las poblaciones a comparar y, en `datos` las observaciones en formato *data frame*.

Ejemplo 5.9 (continuación)

Si no hubiéramos podido validar la normalidad y la homocedasticidad de los datos hubiéramos tenido que utilizar métodos no paramétricos como este test.

Aunque no la utilizaremos, la tabla de rangos de observaciones sería,

Dieta	Rangos						Suma de rangos
A	6	13'5	8'5	7	3		38
B	11'5	15	13'5	3	8'5		51'5
C	10	3	11'5	1	5		30'5

en donde se asigna un rango promedio cuando existen observaciones empatadas.

Con R ejecutamos este test con (1) obteniendo en (2) el valor de estadístico de contraste T y en (3) el p-valor, que sugiere aceptar la hipótesis nula de igualdad de los efectos de las tres dietas.

```
> kruskal.test(peso~dieta,ejemplo) (1)
```

Kruskal-Wallis rank sum test

```
data: peso by dieta
```

```
Kruskal-Wallis chi-squared = 2.2937, df = 2, p-value = 0.3176
```

(2)

(3)

5.9.1. Contraste χ^2 de homogeneidad de varias muestras

Como en las secciones anteriores, este contraste tiene por objeto averiguar si existen o no diferencias significativas entre r poblaciones, de las que se han extraído sendas muestras aleatorias simples. Es válido para comparar dos o más poblaciones.

Es decir, es un contraste semejante —en cuanto a sus propósitos— a los contrastes de análisis de la varianza estudiados anteriormente, aunque con la diferencia de que ahora los datos son frecuencias o recuentos del número de individuos pertenecientes a cada una de las clases en las que se han dividido las poblaciones, y no valores de una variable observable o sus rangos.

Ejemplo 5.11

Con objeto de averiguar si existen o no diferencias significativas entre los hábitos fumadores de tres comunidades, se seleccionó una muestra aleatoria simple de 100 individuos de cada una de las tres comunidades, obteniéndose los siguientes resultados,

Comunidad	fumadores	no fumadores	Total
A	13	87	100
B	17	83	100
C	18	82	100
	48	252	300

¿Pueden considerarse homogéneas las tres poblaciones en cuanto a sus hábitos fumadores?

En general, tendremos s clases (en el ejemplo dos clases, fumadores y no fumadores) en las que se han dividido las r poblaciones, estando clasificadas las r muestras aleatorias extraídas (una de cada población) en una tabla de frecuencias como la anterior en donde cada cruce de fila y columna dará lugar a *celdillas* de frecuencias observadas, n_{ij} , 13, 87, 17,... en el ejemplo.

El propósito de este test es contrastar la hipótesis nula H_0 : *las r poblaciones son homogéneas*, frente a la alternativa de no serlo y el estadístico de contraste es el denominado *estadístico de Pearson* definido como la suma de las *frecuencias observadas* n_{ij} menos las esperadas ne_{ij} si fuera cierta la hipótesis nula anterior, al cuadrado, dividido por la frecuencias esperadas,

$$\lambda = \sum_{\text{celdillas}} \frac{(n_{ij} - ne_{ij})^2}{ne_{ij}}$$

estadístico que sigue, aproximadamente, una distribución χ^2 de Pearson con $(s-1)(r-1)$ grados de libertad, aproximación que será buena si las frecuencias esperadas son, por lo menos, iguales a 5.

Si esto no se cumple, deberemos agrupar clases contiguas —reduciendo adecuadamente los grados de libertad—, o de forma alternativa utilizar el estadístico corregido de Yates.

Contraste de hipótesis

- Aceptar H_0 si $\lambda < \chi^2_{(r-1)(s-1);\alpha}$
- Rechazar H_0 si $\lambda \geq \chi^2_{(r-1)(s-1);\alpha}$

Para ejecutar este test con R la función a utilizar será

`chisq.test(x)`

en donde incluiremos en el primer argumento `x` la matriz de datos.

Ejemplo 5.10 (continuación)

aceptamos la hipótesis nula de homogeneidad de las tres poblaciones en cuanto a sus hábitos fumadores.

Para resolver este ejercicio con R, primero incorporamos los datos en (1) creando la matriz de datos. En (2) y (3) asignamos nombres a las clases que presentan las variables en estudio. Finalmente, en (4) ejecutamos la función `chisq.test` que nos dará la información necesaria sobre el test de homogeneidad de las tres poblaciones.

```
> fuma<-matrix(c(13,17,18,87,83,82),ncol=2) (1)
> colnames(fuma)<-c("fumadores","no fumadores") (2)
> rownames(fuma)<-c("A","B","C") (3)
> chisq.test(fuma) (4)
      Pearson's Chi-squared test
data:  fuma
X-squared = 1.0417, df = 2, p-value = 0.594 (5)
```

En concreto, en (5) obtenemos el valor del estadístico de Pearson, $\lambda = 1'0417$ y del p-valor, $0'594$, suficientemente grande como para concluir con la aceptación de la hipótesis nula de homogeneidad de las tres poblaciones, es decir, con que no existen diferencias significativas entre las tres comunidades en cuanto a sus hábitos fumadores.

Como dijimos, es interesante analizar si las frecuencias esperadas son o no menores que 5 y, para calcularlas debemos ejecutar (6) observamos que las frecuencias esperadas son lo suficientemente grandes como para no requerir agrupar filas y/o columnas contiguas.

```
> chisq.test(fuma)$expected (6)
      fumadores no fumadores
A           16           84
B           16           84
C           16           84
```


Capítulo 6

Modelos de Regresión

6.1. Introducción

En el Ejemplo 1.6 vimos como, a medida que aumentaban los atletas sus horas X de entrenamiento, la marca Y que éstos poseían en 100 metros lisos era menor. De hecho, la Figura 1.6 parece indicarnos que podemos predecir una marca para una horas determinadas de entrenamiento mediante la denominada recta de mínimos cuadrados, también denominada *recta de regresión*, que es la más próxima a la nube de puntos y que en el Capítulo 1 calculamos como

$$y = 15'05908 - 0'04786x.$$

Pero, para toda nube de puntos de consideremos, siempre vamos a poder calcular una recta de regresión que nos permita hacer predicciones de este tipo. La cuestión que nos interesa es saber cuándo estas predicciones son fiables y ése es el propósito principal de la Regresión: analizar, mediante un test de hipótesis, si esta recta es significativa para explicar la variable dependiente Y en función de la independiente X de manera que podamos predecir, por ejemplo, la marca y que conseguiría un atleta que entrenara un tiempo x y, todo esto, con un cierto margen de error que medimos en términos de probabilidades.

Más en concreto, los dos objetivos del Análisis de Regresión que estudiaremos en este capítulo son, analizar si, dados un pares de datos (x_i, y_i) la recta de regresión (o de mínimos cuadrados)

$$y = \beta_0 + \beta_1 x$$

que se obtiene como vimos en el Capítulo 1, es significativa para explicar la variable dependiente Y en función de la variable independiente X y, si esto es así, estimar los *coeficientes de regresión* $\hat{\beta}_0$ y $\hat{\beta}_1$ para hacer predicciones con la ecuación

$$y = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

En realidad, la ordenada en el origen (o *Intercept*) $\widehat{\beta}_0$ se admite que va a estar siempre en la ecuación y no se analiza si es significativa. De hecho, ni siquiera se suele llamar coeficiente de regresión a este parámetro.

6.2. Modelo de la Regresión Lineal Simple

La situación general que se plantea para la *Regresión Lineal Simple* es la de pares de datos (x_i, y_i) procedentes de la observación de dos variables aleatorias, una *independiente* o *covariable*, bajo el control del experimentador, habitualmente representada por X y con valores en el eje de abscisas, y otra denominada *dependiente*, habitualmente representada por Y y con valores en el eje de ordenadas, estando interesados en inferir la existencia o no de una relación lineal entre ambas, de la forma

$$Y = \beta_0 + \beta_1 X + e$$

interpretada ésta en el sentido de que, fijados unos valores x_i , los valores

$$y_{t_i} = \beta_0 + \beta_1 x_i + e_i$$

no son idénticos a los observados y_i debido al error de muestreo e_i .

El Modelo de Regresión Lineal supone que los errores e_i son independientes y con distribución $N(0, \sigma)$, suposiciones que necesitaremos comprobar para que sea válido el test sobre la regresión que explicamos a continuación.

Contraste de la Regresión Lineal Simple

Como hemos dicho anteriormente, en unos casos la recta de regresión podrá ser utilizada para, por ejemplo, hacer predicciones de Y dados unos x concretos y en otros casos no podrá ser utilizada para este propósito porque las predicciones serían desastrosas.

Será la Inferencia Estadística la que deberá ahora validar o no la recta de regresión obtenida, mediante un test de hipótesis en donde la hipótesis nula es $H_0 : X \text{ e } Y \text{ no están relacionadas linealmente}$, (es decir, la recta de regresión no sirve para explicar a la variable dependiente en función de la independiente), y la alternativa $H_1 : X \text{ e } Y \text{ están relacionadas linealmente}$, (es decir, la recta de regresión es útil).

Este test se formaliza formando una *Tabla de Análisis de la Varianza para la Regresión Lineal* en donde se contrasta, repetimos, que todo el modelo es válido o no lo es.

En esta tabla (que es la que da el ordenador), se divide la variación total de los datos en dos fuentes de variación, la *variación explicada por la recta de regresión*, $SSEX$, y la *variación no explicada o residual* $SSNEX$. Si $SSEX$ es grande en relación a $SSNEX$, deberemos rechazar H_0 ; en otro caso aceptarla. El estadístico del test será por tanto, $SSEX/SSNEX$, que hay que estandarizar para que tenga una distribución conocida. En concreto, el estadístico del contraste será

$$F = \frac{SSEX}{SSNEX/(n-2)}$$

que seguirá una distribución F de Snedecor con $(1, n-2)$ grados de libertad.

Contraste de hipótesis

Por lo que antes dijimos, si H_0 es falsa, el estadístico F tenderá a tomar valores grandes, rechazando en ese caso H_0 . Por tanto, el test óptimo de nivel α para contrastar $H_0 : X \text{ e } Y \text{ no están relacionadas linealmente}$, (es decir, la recta de regresión no sirve para explicar a la variable dependiente en función de la independiente), frente a la alternativa, $H_1 : X \text{ e } Y \text{ están relacionadas linealmente}$, (es decir, la recta de regresión es útil), es el siguiente

- | |
|---|
| <ul style="list-style-type: none"> • Se acepta H_0 si $F < F_{1,n-2;\alpha}$ • Se rechaza H_0 si $F \geq F_{1,n-2;\alpha}$ |
|---|

teniendo perfecto sentido el cálculo e interpretación del p-valor del test.

Regresión Lineal con R

La función de R que vamos a utilizar para ejecutar la Regresión Lineal es, primero la función

`lm(modelo)`

incluyendo en el argumento `modelo` la variable dependiente cuantitativa observada, en función de la independiente.

De esta forma obtenemos las estimaciones de los coeficientes de regresión, como ya hicimos en la Sección 1.5.1. El contraste de regresión anterior y la obtención de la tabla de Análisis de la Regresión Lineal se obtienen aplicando la función `anova` al resultado obtenido con la función `lm`.

Ejemplo 6.1

Se midió el contenido de oxígeno, variable Y , a diversas profundidades, variable X , en el lago Worther de Australia, obteniéndose los siguientes datos, en miligramos por litro

X	15	20	30	40	50	60	70
Y	6'5	5'6	5'4	6	4'6	1'4	0'1

Para resolver este ejemplo con R, primero incorporaremos los datos en (1) y (2), obteniendo la recta de regresión, que aquí denominamos **ajus**, al ejecutar (3).

Podemos obtener los estimadores de los coeficientes de regresión ejecutando el objeto creado mediante (4). La recta de regresión ajustada es la que tiene por coeficientes los dados en (5) y que es

$$y = 8'6310 - 0'1081 x$$

Ahora contrastamos la hipótesis nula de que esta recta de regresión no es válida ejecutando (6). El p-valor obtenido en (7) sugiere rechazar la hipótesis nula y concluir que la recta de regresión es válida para explicar la variable dependiente Y en función de la independiente X y, por tano, válida también para hacer predicciones.

```
> x<-c(15,20,30,40,50,60,70) (1)
```

```
> y<-c(6.5,5.6,5.4,6,4.6,1.4,0.1) (2)
```

```
> ajus<-lm(y~x) (3)
```

```
> ajus (4)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

```
(Intercept)          x
      8.6310       -0.1081 (5)
```

```
> anova(ajus) (6)
```

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
x	1	29.4810	29.4810	20.322	0.006352 **
Residuals	5	7.2533	1.4507		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Contraste de hipótesis para β_1

Una forma alternativa al Análisis de la Varianza anterior, para analizar si puede considerarse válida la recta de regresión determinada, es contrastar si

se puede aceptar que es cero o no el coeficiente de regresión β_1 de la ecuación de regresión lineal entre ambas variables.

Si se rechaza la hipótesis nula $H_0 : \beta_1 = 0$ y se acepta la alternativa $H_1 : \beta_1 \neq 0$ la regresión lineal dada por la recta de regresión será aceptable, o en terminología de tests de hipótesis, existe una relación lineal *significativa*, ya que de hecho, el test ha resultado significativo.

Este test alternativo se basa en la distribución en el muestreo del estimador $\widehat{\beta}_1$ y se define en términos de una distribución t de Student.

Si denominamos

$$S_b^2 = \frac{SSNEX/(n-2)}{SSEX/\widehat{\beta}_1^2}$$

el estadístico de contraste

$$t = \frac{\widehat{\beta}_1}{S_b} = \sqrt{\frac{SSEX(n-2)}{SSNEX}}$$

seguirá una distribución t de Student con $n-2$ grados de libertad, por lo que si queremos contrastar $H_0 : \beta_1 = 0$ frente a $H_1 : \beta_1 \neq 0$,

- Se acepta H_0 si $|t| < t_{n-2;\alpha/2}$
- Se rechaza H_0 si $|t| \geq t_{n-2;\alpha/2}$

Para hacer este contraste con R basta con aplicar la función `summary` al resultado obtenido con la función `lm`.

Ejemplo 6.1 (continuación)

Si queremos contrastar la hipótesis nula de ser cero el coeficiente de regresión de X , es decir, $H_0 : \beta_1 = 0$, ejecutamos (8) obteniendo en (9) el p-valor de dicho test, 0'00635, suficientemente pequeño como para rechazar esta hipótesis nula y concluir con que β_1 es significativamente distinto de cero, es decir, que la covariable independiente X es significativa para explicar a la variable dependiente Y mediante la ecuación de la recta de regresión determinada.

```
> summary(ajus) (8)
```

Call:

```
lm(formula = y ~ x)
```

Residuals:

```
1      2      3      4      5      6      7
```

```
-0.50907 -0.86841  0.01289  1.69419  1.37550 -0.74320 -0.96190
```

Coefficients:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.63102     1.07747   8.010  0.00049 ***
x            -0.10813     0.02399  -4.508  0.00635 **
              (9)
```

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 1.204 on 5 degrees of freedom
```

```
Multiple R-Squared:  0.8025,    Adjusted R-squared:  0.7631
```

```
F-statistic: 20.32 on 1 and 5 degrees of freedom,    p-value:    0.006352
```

Hemos obtenido más arriba una tabla ANOVA para analizar la regresión lineal. Esta tabla, no obstante, sólo nos permite contrastar la hipótesis nula de que todo el modelo lineal es adecuado frente a la hipótesis alternativa de no ser todo el modelo lineal ajustado adecuado para explicar los datos que, en el caso de una regresión lineal simple, coincidirá con el test sobre el coeficiente de regresión. No cabe duda de que es más interesante la vía recién estudiada mediante la cual contrastamos la significación de cada covariable que el análisis de todas a la vez.

Por último decir que en la *salida* obtenida al ejecutar `summary` obtenemos, bajo la denominación **Residual standard error**, el estimador de σ , $\sqrt{SSNEX/(n-2)}$ por lo que en el ejemplo anterior, es $\hat{\sigma} = 1'204$.

6.3. Análisis de los residuos

Una de las condiciones necesarias para poder ejecutar los tests anteriores es que, la variable de error e del Modelo Lineal

$$Y = \beta_0 + \beta_1 X + e$$

siga una distribución normal $N(0, \sigma)$. Es decir que, una vez determinada la recta de regresión

$$y_{t_i} = \beta_0 + \beta_1 x_i + e_i$$

los residuos $r_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ deberían de tener una distribución aproximadamente normal $N(0, \sigma)$.

Los residuos los obtenemos ejecutando la función de R `resid` y, el análisis de normalidad lo podemos hacer fácilmente según vimos en la Sección 5.2.

Ejemplo 6.1 (continuación)

Aunque podríamos hacer un análisis gráfico, siempre es mejor ejecutar un test de normalidad, de Kolmogorov-Smirnov, ejecutando (10) o de Shapiro-Wilk ejecutando (11).

```
> ks.test(resid(ajus), "pnorm", 0, 1.204) (10)
```

One-sample Kolmogorov-Smirnov test

```
data: resid(ajus)
D = 0.2352, p-value = 0.7564
alternative hypothesis: two-sided
```

```
> shapiro.test(resid(ajus)) (11)
```

Shapiro-Wilk normality test

```
data: resid(ajus)
W = 0.8219, p-value = 0.06704
```

Aunque ambos tests confirman la normalidad de los residuos, se aprecia de nuevo que el primero es mucho más conservador, especialmente cuando, como pasa aquí, hay pocos datos.

6.4. Modelo de la Regresión Lineal Múltiple

Si en lugar de considerar una sola covariable regresora X , consideramos k covariables independientes tratando de explicar la variable dependiente Y con una ecuación de la forma

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + e$$

estaremos en un caso de *Regresión Lineal Múltiple*.

Al igual que hacíamos con la Regresión Lineal Simple, nuestro objetivo aquí es doble: contrastar qué covariables de las k consideradas puede considerarse significativa para explicar a la variable dependiente Y y, después, estimar los coeficientes de regresión de las covariables que resultaron significativas.

En la Regresión Lineal Múltiple, los residuos también deben de seguir una distribución normal.

En esta doble tarea utilizaremos las mismas funciones de R aunque ya no nos interesa contrastar si toda la ecuación obtenida es o no significativa, sino ejecutar contrastes sobre cada uno de los coeficientes de regresión de forma separada, para estimar finalmente los de las covariables que resultaron significativas.

Ejemplo 6.2

Se consideró que el Número de admisiones previas del paciente, X_1 , y su Edad, X_2 , podrían servir para predecir la Estancia en días, Y , que pasaban en un determinado hospital ciertos enfermos crónicos.

Con dicho propósito se tomó una muestra aleatoria simple de 15 pacientes la cual suministró los siguientes datos

X_1	0	0	0	1	1	1	1	2	2	2	3	3	4	4	5
X_2	21	18	22	24	25	25	26	34	25	38	44	51	39	54	55
Y	15	15	21	28	30	35	40	35	30	45	50	60	45	60	50

Se quiere analizar si alguna o ambas variables independientes X_1, X_2 , pueden servir para explicar a la variable dependiente Y , estimado previamente los coeficientes de regresión de las variables significativas.

El análisis de los coeficientes de regresión lo haremos más adelante, pero ya podemos determinar su estimación con R. Primero incorporamos los datos y, a continuación, se ejecuta (1), obteniendo las estimaciones en (2),

```
> x1<-c(0,0,0,1,1,1,1,2,2,2,3,3,4,4,5)
> x2<-c(21,18,22,24,25,25,26,34,25,38,44,51,39,54,55)
> y<-c(15,15,21,28,30,35,40,35,30,45,50,60,45,60,50)
> hiper<-lm(y~x1+x2)
> hiper
```

(1)

Call:

```
lm(formula = y ~ x1 + x2)
```

Coefficients:

(Intercept)	x1	x2	
2.08572	0.05699	1.05002	(2)

Es decir, el hiperplano de regresión muestral inicialmente propuesto sería

$$y_t = 2'0857 + 0'057 x_1 + 1'05 x_2.$$

Para analizar ahora si ambas covariables son o no significativas ejecutamos (3), observando en (4) los p-valores de los dos tests sobre los coeficientes de regresión, los cuales indican que puede aceptarse la hipótesis nula de ser cero el coeficiente de regresión de X_1 , debiendo eliminar esta variable del modelo, pero que la covariable X_2 sí es significativa.

```
> summary(hiper)
```

(3)

Call:

```
lm(formula = Y ~ x1 + x2)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.122	-3.543	1.542	2.317	10.557

Coefficients:


```

              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.08572     6.73931   0.309  0.76226
x1           0.05699     2.61310   0.022  0.98296
x2           1.05002     0.32621   3.219  0.00737 **
                                     (4)
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.059 on 12 degrees of freedom
Multiple R-Squared:  0.8503,    Adjusted R-squared:  0.8254
F-statistic: 34.08 on 2 and 12 DF,  p-value: 1.125e-05

```

Con objeto de completar el ejemplo, ejecutamos (5) y (6), obteniendo en (7) los coeficientes de la recta de regresión lineal ajustada, cuyo p-valor asociado, (8), confirma que la Edad del paciente, X_2 , es significativa (ahora aún más) para explicar a la variable dependiente, Estancia en días en el hospital.

```

> hiper2<-lm(Y ~ x2)                                     (5)
> summary(hiper2)                                         (6)

```

```

Call:
lm(formula = Y ~ x2)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-10.089  -3.561   1.534   2.345  10.552

```

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.977       4.373   0.452   0.659
x2             1.057       0.123   8.593 1.01e-06 ***
                                     (7)

```

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.821 on 13 degrees of freedom
Multiple R-Squared:  0.8503,    Adjusted R-squared:  0.8388
F-statistic: 73.84 on 1 and 13 DF,  p-value: 1.014e-06

```

La recta de regresión finalmente ajustada será por tanto,

$$y_t = 1'977 + 1'057 x_2$$

la cual permite predecir, por ejemplo, un paciente de 60 años que ingrese en el hospital en estudio es muy probable que esté en él,

$$y_t = 1'977 + 1'057 \cdot 60 = 65'397$$

días.

6.5. Otros Modelos Lineales

Con la Regresión Lineal Múltiple (y Simple) analizamos si k covariables independientes $X_1, \dots, X_k$ son significativas para explicar a la variable dependiente Y mediante una ecuación de la forma

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + e.$$

Aunque no lo hemos dicho, tanto las k covariables independientes como la dependiente deben de ser de tipo cuantitativo. Si las k covariables independientes fueran todas ellas de tipo cualitativo estaríamos en un caso de Análisis de la Varianza, como el que estudiamos en la Sección 5.8, en donde las k variables serían los r tratamientos considerados, pero con una salvedad que consiste en que, para expresar un Análisis de la Varianza como Modelo Lineal, debemos emplear tantas covariables de tipo indicador o dummy (con sólo dos valores 0 y 1) $X_1, X_2, \dots$ como clases o “valores” tenga el Tratamiento en estudio, menos una. Es decir, si queremos expresar el Ejemplo 5.9 como Modelo Lineal podemos escribir el Aumento de Peso Y en función de dos covariables de la forma

$$Y = \beta_0 + \beta_1 X_A + \beta_2 X_B + e$$

siendo (X_A, X_B) dos variables que tomarán el valor $(1, 0)$ cuando queramos predecir aumentos de peso en ratones sometidos a la dieta A , que tomarán el valor $(0, 1)$ cuando queramos predecir aumentos de peso en ratones sometidos a la dieta B y que tomarán el valor $(0, 0)$ cuando queramos predecir aumentos de peso en ratones sometidos a la dieta C .

Se hablará de *Análisis de la Covarianza* cuando entre las k covariables independientes algunas sean de tipo cualitativo y otras son de tipo cuantitativo. Estos tres tipos de modelos reciben el nombre común de *Modelos Lineales* porque la variable dependiente Y se expresa como una *función lineal* de los parámetros $\beta_0, \beta_1, \dots, \beta_k$.

Los Modelos Lineales se ajustan con la función lm de R y el propósito es siempre el mismo: primero, analizar qué covariables independientes $X_1, \dots, X_k$ son significativas para explicar a la variable dependiente Y y, segundo, estimar los coeficientes de regresión de las que resultaron significativas con objeto de hacer predicciones. Además, siempre debemos analizar si los residuos siguen una distribución normal.

Si la expresión que relaciona a las covariables independientes y la dependiente no fuera lineal, se hablaría de *Modelos no Lineales*, pero es más habitual generalizar los Modelos Lineales considerando lo que se denomina *Modelos Lineales Generalizados* en donde se considera como variable dependiente Y , en lugar de una variable del tipo Peso o Talla como en los Modelos Lineales, una

variable dicotómica que sólo puede tomar dos valores 0 ó 1 correspondientes a *éxito-fracaso*, es decir, *ocurrencia-no ocurrencia* del algún suceso del tipo supervivencia o fallecimiento de pacientes en estudio. Este tipo de modelos recibe el nombre de *Regresión Logística*.

Si la variable dependiente Y pudiera tomar valores del tipo $0, 1, 2, \dots$, como por ejemplo número de supervivientes a una determinada enfermedad, el modelo se denominaría de *Regresión Poisson*.

Estos dos últimos modelos expresan la relación entre la variable dependiente Y y las k covariables independientes de forma algo diferente, por ejemplo mediante logaritmos y, junto con los Modelos Lineales, forman lo que se denominan *Modelos Lineales Generalizados*, los cuales se ajustan con la función `glm` y en donde el propósito es, de nuevo, analizar qué covariables independientes (cualitativas y cuantitativas) son significativas para explicar a la variable dependiente Y y estimar los coeficientes de regresión de las que resultaron significativas. Los residuos de todos estos modelos deben de tener una distribución normal. Los lectores interesados en este tipo de modelos, pueden estudiarlos en el texto de este autor, *Métodos Avanzados de Estadística Aplicada. Técnicas Avanzadas*.

Los Modelos Lineales también se pueden extender permitiendo a las covariables independientes X_i una expresión más general que la anterior mediante unas funciones h_i , aunque manteniendo la linealidad del modelo, de la forma

$$Y = h_0 + h_1(X_1) + \dots + h_k(X_k) + e.$$

La incorporación de las funciones h_i hace que el modelo sea más flexible y capaz de adaptarse a datos más complejos que no muestren una estricta linealidad en las covariables. No obstante, los modelos aditivos tienen que verificar todas las suposiciones que exigíamos a los modelos de regresión lineal como la normalidad de los residuos y la homocedasticidad. Este modelo se denominan *Modelos Aditivos*.

Si generalizamos los Modelos Aditivos de la misma manera que los Modelos Lineales Generalizados GLM generalizaban los Modelos Lineales tendremos los denominados *Modelos Aditivos Generalizados* GAM que constituyen la clase de modelos más general, aunque el propósito sigue siendo el mismo: analizar qué covariables independientes son significativas para explicar a la variable dependiente y estimar los coeficientes de regresión de las que resultaron significativas. Aquellos lectores interesados en este tipo de modelos y en los GLM, pueden leer el texto de este autor, *Técnicas Actuales de Estadística Aplicada*.

Capítulo 7

Bibliografía

- Affi, A.A. y Clark, V. (1990). *Computer-aided Multivariate Analysis*. Belmont, California: Lifetime Learning Publications.
- De Moivre, A. (1733). Approximatio ad Summam Terminorum Binomii $(a + b)^n$ in Seriem expansi. Opúsculo en Latín del 12 de Noviembre de 1733.
- Dolkart, R.E., Halperin, B. y Perlman, J. (1971). Comparison of antibody responses in normal and alloxan diabetic mice. *Diabetes*, **20**, 162-167.
- García Pérez, A. (1993). *Estadística Aplicada con BMDP*. UNED. Colección Educación Permanente.
- García Pérez, A. (1993). *Estadística Aplicada con SAS*. UNED. Colección Educación Permanente.
- García Pérez, A. (1998). *Fórmulas y Tablas Estadísticas*. UNED. Colección Adenda.
- García Pérez, A. (1998). *Problemas Resueltos de Estadística Básica*. UNED. Colección Educación Permanente.
- García Pérez, A. (2005). *Métodos Avanzados de Estadística Aplicada. Técnicas Avanzadas*. UNED. Colección Educación Permanente.
- García Pérez, A. (2005). *Métodos Avanzados de Estadística Aplicada. Métodos Robustos y de Remuestreo*. UNED. Colección Educación Permanente.
- García Pérez, A. (2008). *Estadística Aplicada: Conceptos Básicos*. Segunda edición. UNED. Colección: Educación Permanente.
- García Pérez, A. (2008). *Ejercicios de Estadística Aplicada*. UNED. Colección: Cuadernos de la UNED.
- García Pérez, A. (2008). *Estadística Aplicada con R*. Editorial UNED. Colección Varia.
- García Pérez, A. (2010). *Estadística Básica con R*. Editorial UNED. Colección Grado.
- García Pérez, A. (2015). *Técnicas Actuales de Estadística Aplicada*. Editorial UNED. En prensa.
- Gauss, C. F. (1809). *Theoria motus corporum coelestium in sectionis conicis solem ambientum*, Hamburgo.
- Johnson, G.A. (1973). *Local Exchange and Early State Development in Southwestern Iran*. University of Michigan Museum of Anthropology, Anthropological Papers n. 51. University of Michigan, Ann Arbor.

- Laplace, P-S de (1814). *Essai Philosophique sur les probabilités*. (Existe traducción: *Ensayo filosófico sobre las probabilidades*, Alianza.)
- Student (1908). The probable error of a mean. *Biometrika*, **6**, 1-25.
- van Oost, B.A., Veldhayzen, B., Timmermans, A.P.M. y Sixma, J.J.(1983). Increased urinary β -thromboglobulin excretion in diabetes assayed with a modified RIA kit-technique. *Thrombosis and Haemostasis*, **9**, 18-20.
- Weiner, B. (1977). *Discovering Psychology*, Chicago: Science Research Association, 97.

En esta publicación se estudian los principales conceptos de la Estadística Aplicada, y va dirigido a los lectores que no tienen ningún conocimiento previo de dicha materia. Es, por tanto, un libro iniciático en dicha área, la cual cada día tiene mayor importancia en la sociedad. Como hoy en día es muy conveniente la utilización del ordenador, el texto se ha escrito ilustrando la exposición de los conceptos y métodos estadísticos con la ayuda del paquete estadístico R, el mejor y más utilizado, y que, además, es gratuito.

Alfonso García Pérez es, desde 1996, catedrático del área Estadística e Investigación Operativa en la UNED. En 1983 fue adjunto de Bioestadística y, en 1984, adjunto de Estadística Matemática y Cálculo de Probabilidades en la Universidad Autónoma de Madrid. Además de esta intensa actividad docente, tiene publicados 16 libros (14 de ellos en la UNED) y más de 40 artículos de investigación en revistas internacionales de prestigio, y presentado más de 56 comunicaciones en congresos de investigación nacionales e internacionales.



UNED

Editorial

ciencias



0105008CT01A01